TELECOMMUNICATIONS

VOLUME 5

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

Editor John G. Proakis

Editorial Board Rene Cruz University of California at San Diego Gerd Keiser Consultant Allen Levesque Consultant Larry Milstein University of California at San Diego Zoran Zvonar Analog Devices *Editorial Staff* Vice President, STM Books: Janet Bailey Sponsoring Editor: George J. Telecki Assistant Editor: Cassie Craig

Production Staff Director, Book Production and Manufacturing: Camille P. Carter Managing Editor: Shirley Thomas Illustration Manager: Dean Gonzalez

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 5

John G. Proakis Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at http://www.mrw.interscience.wiley.com/eot

Copyright © 2003 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permreq@wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor. p. cm.
includes index. ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G. TK5102_W55 2002

621.382′03 — dc21

2002014432

Printed in the United States of America

 $10\ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2\ 1$

SPATIOTEMPORAL SIGNAL PROCESSING IN WIRELESS COMMUNICATIONS

DIMITRIOS HATZINAKOS RYAN A. PACHECO University of Toronto Toronto, Ontario, Canada

1. INTRODUCTION

Wireless channels are severely limited by fading due to multipath propagation. When fading is "flat," the receiver experiences fast or slow variations of the received signal amplitude, which makes detection of the information symbols difficult. Fading can also be frequency-selective, in which case in addition to signal variations we experience intersymbol interference. An effective strategy to deal with multipath propagation and fading has been the utilization of multiple antennas at the receiver and or at the transmitter. Beamforming with multiple antennas closely spaced together has been traditionally used for directive transmission and rejection of multipath components. However, this approach requires line-of-sight communications and the formation of steerable multiple beams in the case of mobile multiuser communications. Another approach is diversity, which is based on the weighted combination of multiple signal copies that are subject to independent fading. Therefore, two adjacent receive or transmit antennas must be apart by several wavelengths so that signals transmitted or received from different antennas are sufficiently uncorrelated. The latter approach is considered in this article.

The information-theoretic capacity of multiple antenna systems and data communication techniques that exploit the potential of spatial diversity have been studied [3,4], and dramatic increases in supportable data rates are expected. In such systems, the capacity can theoretically increase by a factor up to the number of transmit and receive antennas in the array. Spacetime diversity and coding techniques, such as trellis spacetime codes, block spacetime codes, and variations of the very popular BLAST (Bell Laboratories layered spacetime) architecture, among others [5,7–9] are currently under investigation. However, there is a need for more thorough investigation of the potential and limits of such technologies and the development of efficient low complexity systems. In the first half of this article, we provide a general treatment of the spacetime signal processing scenario and spacetime diversity and coding frameworks based on block spacetime codes and discuss potential benefits, requirements, and limitations.

A common assumption in existing spacetime coding literature is the knowledge or availability of a good estimate for the multipath channel [3,7]. In the absence of channel knowledge, the capacity gains to be achieved depend on the particular characteristics (e.g., coherence time) of the channel. Various schemes, consisting of an antenna array and a spatio-temporal processor at the base station (or at the mobile unit), have been proposed to mitigate the effects of intersymbol interference (ISI), and to reject cochannel interference (CCI) or multipleaccess interference (MAI) induced by simultaneous intraor intercell users [5,17]. To cope with the time-varying and fading nature of wireless channels, data are transmitted in bursts, and attached to each burst is a training sequence of short duration that is used to help the receiver recover the unknown parameters of the communication channel [10]. In many cases unsupervised (i.e., blind) techniques [12,21], which require no training data, as well as semiblind approaches, which provide a synergistic treatment of training and blind principles, are employed to address the channel estimation and/or the direct channel equalization problems [1,11,14]. In the second half of the article, a semiblind channel equalization approach will be presented and integrated with spacetime diversity and coding to form a realistic scenario for detection and estimation.

The aim of the article is by no means to provide an exhaustive coverage but rather to illustrate the principle and potential and to suggest a realistic approach for implementation of spacetime wireless transceivers, and also to provide a motivation for exploring the fast-growing literature and technological developments in the area.

2. THE DISCRETE-CHANNEL MODEL

The basic structure of a wireless transceiver employing spacetime diversity and coding (multiple-input multipleoutput system) is depicted in Fig. 1.

The received equivalent discrete-time signal at the receiver antenna a, a = 1, ..., A, after demodulation, filtering, and oversampling (i.e., taking N samples per symbol period) can be written as [1]

$$r_a[n] = \sum_{m=1}^{M} \sum_{k=0}^{N_b - 1} b_m[k] \cdot g_{ma}[n - kN] + v_a[n]$$
(1)

where N_b is the length of the transmitted burst of data and the $\{b_m[k]\}_{k=0}^{N_b-1}$ denote the data symbols transmitted from the antenna $m, m = 1, \ldots, M$. When spacetime coding is employed in the transmitter, we may assume that M



Figure 1. Single-user communication using multiple transmit and receive antennas.

complex symbols are transmitted simultaneously from the M transmit antennas. In this case, we may write [3]

$$\mathbf{b}[k] = \sum_{l=1}^{M} \mathbf{q}_l^r[k] \cdot s_l^r + \sum_{l=1}^{M} j \mathbf{q}_l^j[k] \cdot s_l^j$$
(2)

where $\mathbf{b}(k) = [b_1(k), \dots, b_M(k)]^T$ is the transmitted data vector at time instant k, $s_l = s_l^r + j s_l^j$, $l = 1, \dots, M$ are the transmitted symbols and $\mathbf{q}_l^r(k)$ and $\mathbf{q}_l^j(k)$ are the $M \times 1$ transmit antenna weight vectors for the real and imaginary part of the symbol l, respectively, at time instant k. The $g_{ma}[n], n = 0, \dots, LN - 1$ accounts for the transmit and receive filters and the multipath channel between transmit antenna m and receive antenna a. Without loss of generality we will assume that all the channels between transmit and receive antennas are of the same length LN - 1 samples and are time-invariant during the transmission of a burst of N_b data symbols. Finally $v_a[n]$ is circular complex white Gaussian noise, uncorrelated with $b_m[n] \forall m$, with zero mean and variance σ^2 .

Assume that we are interested in recovering the *m*th transmitted symbol at time instant *k*. To do this we first collect *N* samples from all antenna elements to form the vector $\mathbf{r}(k)$, which takes the following form:

$$\mathbf{r}(k) = [\mathbf{G}(L-1), \dots, \mathbf{G}(0)]$$
$$\cdot [\mathbf{b}(k-L+1)^T, \dots, \mathbf{b}(k)^T]^T + \mathbf{v}(k)$$
(3)

where

$$\mathbf{r}(k) = \begin{bmatrix} r_{1}[kN] \\ \vdots \\ r_{1}[(k+1)N-1] \\ \vdots \\ r_{A}[kN] \\ \vdots \\ r_{A}[(k+1)N-1] \end{bmatrix}_{AN \times 1}$$

$$\mathbf{G}(l) = \begin{bmatrix} g_{11}[lN] & \cdots & g_{M1}[lN] \\ \vdots & & \vdots \\ g_{11}[(l+1)N-1] & \cdots & g_{M1}[(l+1)N-1] \\ \vdots & & \vdots \\ g_{11}[(l+1)N-1] & \cdots & g_{M1}[(l+1)N-1] \\ \vdots & & \vdots \\ g_{1A}[(l+1)N-1] & \cdots & g_{MA}[(lN] \\ \vdots \\ g_{1A}[(l+1)N-1] & \cdots & g_{MA}[(l+1)N-1] \end{bmatrix}_{AN \times M}$$

$$\mathbf{b}(k) = \begin{bmatrix} b_{1}[k] \\ \vdots \\ b_{M}[k] \end{bmatrix}_{M \times 1} \mathbf{v}(k) = \begin{bmatrix} v_{1}[kN] \\ \vdots \\ v_{1}[(k+1)N-1] \\ \vdots \\ v_{A}[kN] \\ \vdots \\ v_{A}[(k+1)N-1] \end{bmatrix}_{AN \times 1}$$

In general, the simultaneous transmission of M symbols will span μ received vectors. Thus, it may be necessary to process more than one received vector at a time in order to estimate the *m*th symbol. Then, assuming a linear receiver structure and in the absence of noise, the recovery of the real information symbol s_m^q , where q = r or q = j, at time instant k, requires that the $A \times 1$ weight vectors $\mathbf{W}_{i,m}^q$ of the spacetime equalizer at $i = k, \ldots, k + \mu - 1$ satisfy the relation

$$\sum_{i=k}^{k+\mu-1} \mathbf{W}_{i,m}^{q^H} \cdot \mathbf{r}(i) = s_m^q [k-d]$$
(4)

where d is an arbitrary delay. Throughout the article the symbols *, T, and H denote conjugate, transpose, and conjugate transpose operations, respectively.

3. SPACETIME DIVERSITY AND FLAT-FADING CHANNELS

Here we are going to examine the benefits that can be expected by employing both transmit and receive diversities. We follow a procedure based on an excellent analysis [3] of this problem. To simplify the mathematical formulas, we assume a flat-fading channel described by the constant matrix **G** and that N = 1, with no oversampling. Then, by placing L = 1, N = 1, and $\mathbf{G}(0) = \mathbf{G}$ into the relation of Section 2, we obtain the following simplified relation between the $A \times 1$ receive signal vector $\mathbf{r}(k)$ and the $M \times 1$ transmit data symbol vector $\mathbf{b}(k)$:

$$\mathbf{r}(k) = \mathbf{G} \cdot \mathbf{b}(k) + \mathbf{v}(k) \tag{5}$$

where

$$\mathbf{r}(k) = \begin{bmatrix} r_1[k] \\ \vdots \\ r_A[k] \end{bmatrix}_{A \times 1} \quad \mathbf{G} = \begin{bmatrix} g_{11} & \cdots & g_{M1} \\ \vdots & & \vdots \\ g_{1A} & \cdots & g_{MA} \end{bmatrix}_{A \times M}$$
$$\mathbf{b}(k) = \begin{bmatrix} b_1[k] \\ \vdots \\ b_M[k] \end{bmatrix}_{M \times 1} \quad \mathbf{v}(k) = \begin{bmatrix} v_1[k] \\ \vdots \\ v_A[k] \end{bmatrix}_{AN \times 1}$$

The transmitted data vector at time instant k is given by Eq. (2). Then, according to Eq. (4), the output of the linear spatiotemporal receiver filter (detector) for the real symbol $s_m^q, q = r$ or q = j, m = 1, ..., M is

$$\begin{split} \hat{s}_{m}^{q} &= Real \left\{ \sum_{i=k}^{k+\mu-1} \mathbf{W}_{i,m}^{q^{H}} \cdot \mathbf{r}(i) \right\} \\ &= Real \left\{ \sum_{i=k}^{\mu+k-1} \mathbf{W}_{i,m}^{q^{H}} \mathbf{G} \sum_{l=1}^{M} \mathbf{q}_{l}^{r}(i) s_{l}^{r} \right. \\ &+ \left. \sum_{i=k}^{\mu+k-1} \mathbf{W}_{i,m}^{q^{H}} \mathbf{G} \sum_{l=1}^{M} j \mathbf{q}_{l}^{j}(i) s_{l}^{j} + \left. \sum_{i=k}^{\mu+k-1} \mathbf{W}_{i,m}^{q^{H}} \mathbf{v}(i) \right\} \quad (6) \end{split}$$

It can be shown [3,6] that the maximum signal to noise ratio (SNR) for this linear spatiotemporal receiver is achieved when (within a scale term)

$$\mathbf{W}_{i,m} = \mathbf{G}\mathbf{q}_m^r(i) \qquad \text{or} \qquad \mathbf{W}_{i,m} = \mathbf{G}_j \mathbf{q}_m^j(i) \tag{7}$$

for s_m^r and s_m^i , respectively. By substituting these values in Eq. (6), we determine the detector output \hat{s}_m^r for s_m^r , where $m = 1, \ldots, M$:

$$\begin{split} \hat{s}_{m}^{r} &= Real \left\{ \sum_{i=k}^{\mu+k-1} \mathbf{q}_{m}^{r^{H}}(i) \mathbf{G}^{H} \mathbf{G} \sum_{l=1}^{M} \mathbf{q}_{l}^{r}(i) s_{l}^{r} \right\} \\ &+ Real \left\{ \sum_{i=k}^{\mu+k-1} \mathbf{q}_{m}^{r^{H}}(i) \mathbf{G}^{H} \mathbf{G} \sum_{l=1}^{M} j \mathbf{q}_{l}^{j}(i) s_{l}^{j} \right\} \\ &+ Real \left\{ \sum_{i=k}^{\mu+k-1} \mathbf{q}_{m}^{r^{H}}(i) \mathbf{G}^{H} \mathbf{v}(i) \right\} \\ &= Real \left\{ \sum_{l=1}^{M} trace(\mathbf{G} \mathbf{Q}_{l}^{r} \mathbf{Q}_{m}^{r^{H}} \mathbf{G}^{H}) s_{l}^{r} \right\} \\ &+ Real \left\{ \sum_{l=1}^{M} j trace(\mathbf{G} \mathbf{Q}_{l}^{j} \mathbf{Q}_{m}^{r^{H}} \mathbf{G}^{H}) s_{l}^{j} \right\} \\ &+ Real \left\{ \sum_{l=1}^{M} j trace(\mathbf{G} \mathbf{Q}_{l}^{j} \mathbf{Q}_{m}^{r^{H}} \mathbf{G}^{H}) s_{l}^{j} \right\} \end{split}$$
(8)

where $\mathbf{Q}_m^r = [\mathbf{q}_m^r(k), \ldots, \mathbf{q}_m^r(k+\mu-1)]_{M \times \mu}$ and $\mathbf{Q}_m^j = [\mathbf{q}_m^j(k), \ldots, \mathbf{q}_m^j(k+\mu-1)]_{M \times \mu}$ are the transmitter weight matrices for symbols s_m^r and s_m^j , respectively. Similarly, we find the detector output \hat{s}_m^j for s_m^j , where $m = 1, \ldots, M$:

$$\hat{s}_{m}^{j} = -Real \left\{ \sum_{l=1}^{M} trace(\mathbf{G}\mathbf{Q}_{l}^{j}\mathbf{Q}_{m}^{j^{H}}\mathbf{G}^{H})s_{l}^{j} \right\}$$

$$+ Real \left\{ \sum_{l=1}^{M} jtrace(\mathbf{G}\mathbf{Q}_{l}^{r}\mathbf{Q}_{m}^{j^{H}}\mathbf{G}^{H})s_{l}^{r} \right\}$$

$$+ Real \left\{ \sum_{i=k}^{\mu+k-1} \mathbf{q}_{m}^{j^{H}}(i)\mathbf{G}^{H}\mathbf{v}(i) \right\}$$

$$(9)$$

By properly choosing the weight matrices $\mathbf{Q}_m^r, \mathbf{Q}_m^j, m = 1, \ldots, M$, the interference from other symbols can be significantly reduced or completely eliminated and the output SNR maximized. In this case, a transmit diversity gain up to the order M and a transmit-receive diversity gain of order up to AM is possible. Another way of looking at the above procedure is that of mapping the set of symbols $\{s_l\}$ to the set of vectors $\{\mathbf{b}(k)\}$, that is

$$[s_1 s_2 \cdots s_M]_{1 \times M} \to [\mathbf{b}(k) \mathbf{b}(k+1) \cdots \mathbf{b}(k+\mu-1)]_{M \times \mu} \quad (10)$$

which is a form of spacetime block code. Thus, the objective is to design efficient such codes. It can be shown that in general this is possible only if $\mu \ge M$ [3–5]. Next we provide a few examples to demonstrate some important points.

Example 1. Let us consider the transmission of two complex symbols s_1, s_2 from a transmitter with two antennas by using the above mentioned process (M = 2) and let $\mu = 2$. First let us consider the following choice of weight matrices

$$\mathbf{Q}_1^r = [\mathbf{q} \ \mathbf{0}]\mathbf{Q}_2^r = [\mathbf{0} \ \mathbf{q}]\mathbf{Q}_1^j = [\mathbf{q} \ \mathbf{0}]\mathbf{Q}_2^j = [\mathbf{0} \ \mathbf{q}]$$

where ${\boldsymbol{q}}$ is a 2×1 weight vector. The transmitted complex vectors are

$$\mathbf{b}(k) = [\mathbf{q}s_1], \quad \mathbf{b}(k+1) = [\mathbf{q}s_2]$$

This is clearly the situation of Space-only transmission. Then, since

$$\mathbf{Q}_l^x \mathbf{Q}_m^{x^H} = \mathbf{0}, \quad x = r, j, \quad m \neq l$$

interference from different symbols is eliminated. Furthermore since

$$\mathbf{Q}_m^x \mathbf{Q}_m^{y^{-1}}, \quad m = 1, 2$$

are Hermitian, the terms

$$Trace(\mathbf{G}\mathbf{Q}_m^x \mathbf{Q}_m^{y^H} \mathbf{G}^H), \quad m = 1, 2$$

are real. Thus, equations (8) and (9) are reduced to

$$D_m^x = \pm trace(\mathbf{G}\mathbf{q}\mathbf{q}_m^H \mathbf{G}^H) s_m^x + \mathbf{q}^H \mathbf{G}^H \mathbf{v}$$
$$x = r, j \quad m = 1, 2 \tag{11}$$

and the corresponding SNR per real symbol is easily found to be

$$SNR = \frac{E_s}{\sigma^2} trace(\mathbf{Gqq}^H \mathbf{G}^H)$$

where E_s is the energy of the real symbol and σ^2 the variance of the white noise.

Example 2. Let again M = 2 and $\mu = 2$ but this time choose the weight matrices as follows [3-5]:

$$\begin{aligned} \mathbf{Q}_1^r &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Q}_2^r = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \mathbf{Q}_1^j = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \\ \mathbf{Q}_2^j &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{aligned}$$

and the transmitted complex vectors are

$$\mathbf{b}(k) = \begin{bmatrix} s_1 \\ -s_2^* \end{bmatrix}, \quad \mathbf{b}(k+1) = \begin{bmatrix} s_2 \\ s_1^* \end{bmatrix}$$

In this case of *spacetime* processing it is easy to show by following arguments similar to those in example 1 that all interference is eliminated and since

$$\mathbf{Q}_m^x \mathbf{Q}_m^{x^H} = \mathbf{I}_{2 \times 2}, \quad m = 1, 2$$

are identity matrices

$$D_m^x = \pm trace(\mathbf{G}\mathbf{G}^H)s_m^x + \sum_{i=k}^{k+1} \mathbf{q}_m^{x^H}(i)\mathbf{G}^H(0)\mathbf{v}(i)$$
$$x = r, j \quad m = 1, 2$$
(12)

and the corresponding SNR per real symbol is found to be

$$\text{SNR} = \frac{E_s}{\sigma^2} trace(\mathbf{G}\mathbf{G}^H) = \frac{E_s}{\sigma^2} \sum_{i=1}^2 \sum_{j=1}^2 |g_j^i|^2$$

Example 3. Consider now the case of M = 1, A > 1 corresponding to pure receive diversity. In this case the matrix G reduces to a vector $A \times 1$. Similarly, the transmit weight matrices reduce to scalars and after some calculations we find that the corresponding maximum SNR per real symbol is

$$\mathrm{SNR} = rac{E_s}{\sigma^2} \sum_{i=1}^A |g_1^i|^2$$

which corresponds to the case of *Maximal Ratio Combining*.

From these examples, we conclude the following:

- 1. In the case of space-only processing (Example 1), it is important that both the transmitter and the receiver know the channel, that is, the matrix \mathbf{G} , to maximize the output SNR. This is achieved when the transmit weight vector \mathbf{q} is equal to the eigenvector of \mathbf{G} corresponding to the largest eigenvalue.
- 2. On the other hand, in the case of spacetime processing of Example 2, the transmitter does not need to know the channel. Obviously the receiver still requires knowledge of the channel for decoding.
- 3. Examples 2 and 3 indicate that the spacetime diversity gain with M transmit and A receive antennas is of the order $M \cdot A$. Thus, in Example 3, in order to achieve similar gains as in Example 2, we must choose A = 4. Similar arguments can be made for the case of transmit only diversity where M > 1, A = 1.
- 4. Spacetime diversity exploits the propagation environment itself to improve the performance. The *richer* the multipath channel (matrix G), the higher the diversity gain is.

4. SPACETIME DIVERSITY AND SEMIBLIND CHANNEL EQUALIZATION

In multiaccess communications, channel resources are allocated in ways that require either strict cooperation among users, such as in time- or frequency-division multiple access (TDMA or FDMA), or not, as in codeor space-division multiple access (DS-CDMA or SDMA). DS-CDMA and SDMA systems generally require more complex receivers than do TDMA or FDMA since the received data may contain interference from other users (multiuser interference, also called *multiple-access interference* (MAI)). The amount of interference that one user contributes to another is dependent on the orthogonality between their received signals. In DS-CDMA systems, orthogonality may exist on the transmitting end by assigning orthogonal spreading codes to users but will not exist on the receiver end because of propagation effects (multipath propagation in the case of wireless channels). The end result is that correlation-type receivers are no longer viable and some method of multiuser detection is needed in order to deal effectively with MAI. Many different multiuser detection strategies have been proposed over the past decade from prohibitively complex [20] to extremely simple [13].

In addition to MAI there is also intersymbol interference (ISI), or *self-interference*, that is also caused by multipath propagation. ISI is negligible in low-rate DS-CDMA applications but is becoming more of an issue in third-generation systems (where the symbol duration is on the order of the multipath delay spread [10]). Consequently, high-rate DS-CDMA systems suffer from MAI and ISI. Thus, receivers for such systems need to perform equalization and multiuser detection.

When MAI and ISI are present, an analytic treatment of the spacetime diversity and coding scenario is not as straightforward as with flat-fading single-user channels. In any case, the unknown wireless channel must be estimated at the receiver or directly equalized for cancellation of the interference and detection of the desired data symbols. In this section, we complement the spacetime diversity and coding principles described in the previous section with a spacetime semiblind channel equalization technique and propose a realistic and effective wireless communications transceiver.

4.1. Revised Signal Model

A K-user DS-CDMA system is considered. Each user is assigned a unique spreading code with a processing gain of N. In addition, each user transmits from an *M*-element antenna array to an antenna array with A elements. The block diagram of such a transceiver is depicted in Fig. 2. Let $b_m^k[n]$ be the *n*th symbol for the *k*th user from the *m*th transmit antenna. We assume a 4-QAM modulation scheme $b_m^k[n] \in \{e^{\pm j(\pi/4)}, e^{\pm j(3\pi/4)}\}$. Summing over all transmit antennas for all users, we can rewrite the received signal at the *a*th receive antenna



Figure 2. kth-user transceiver using 2 transmit and A receive antennas including spacetime coding and channel equalization.

is as

$$r_{a}[n] = \sum_{k=1}^{K} \sum_{m1}^{M} \sum_{k=0}^{N_{b}-1} b_{m}^{k}[n_{b}]g_{ma}^{k}[n-Nn_{b}] + v_{a}[n]$$
(13)

where now $g_{ma}^{k}[n]$ represents the combined impulse response of channel plus spreading code for the *k*th user from the *m*th transmit antenna to the *a*th receive antenna:

$$g_{ma}^{k}[n] = \sum_{l=0}^{L_{k}-1} \beta_{ma} k[l] c_{k}[n-l]$$
(14)

In this equation $\{\beta_{ma}^{k}[l]\}_{l=0}^{L_{k}-1}$ represents the frequencyselective multipath channel for the given user. The $c_{k}[n]$ is the normalized spreading code for the *k*th user: $c_{k}[n] \in \{\pm 1\}/\sqrt{N}$. Once again, it is assumed that the fading coefficients do not change during the transmission of N_{b} data symbols and that $\max(L_{k}-1) \leq (L-1)N$ for some integer *L*.

Collecting the N chips corresponding to the n_b th transmitted symbol, the received vector is written as

$$\mathbf{r}_a(n_b) = \mathbf{G}_a \cdot \mathbf{b}_{n_b} + \mathbf{v}_a(n_b) \tag{15}$$

where

$$\mathbf{r}_{a}(l) = [r_{a}[lN], \dots, r_{a}[(l+1)N-1]]^{T}$$

$$\mathbf{G}_{a} = [\mathbf{G}_{a}(L-1), \dots, \mathbf{G}_{a}(0)]$$

$$\mathbf{G}_{a}(l) = [\mathbf{g}_{1a}^{1}(l), \mathbf{g}_{2a}^{1}(l), \dots, \mathbf{g}_{Ma}^{K}(l)]$$

$$\mathbf{g}_{ia}^{k}(l) = [g_{ia}^{k}[lN], \dots, g_{ia}^{k}[(l+1)N-1]]^{T}$$

$$\mathbf{b}_{nb} = [\mathbf{b}(n_{b}-L+1)^{T}, \dots, \mathbf{b}(n_{b})^{T}]^{T}$$

$$\mathbf{b}(l) = [b_{1}^{1}[l], b_{2}^{1}[l], \dots, b_{M}^{K}[l]]^{T}$$

$$\mathbf{v}_{a}(l) = [v_{a}[lN], \dots, v_{a}[(l+1)N-1]]^{T}$$

4.2. Interference Suppression Strategies

We now fix the number of transmit antennas per user (M) at two and apply the block code of Example 2 (this is similar to the Alamouti scheme [7]). In this case, symbols for the *k*th user, $\{b_k[n]\}_{n=0}^{N_b-1}$, are mapped as follows for each pair of symbols ($\{b_k[0], b_k[1]\}, \{b_k[2], b_k[3]\}, \ldots$):

$$b_1^k[0] = \frac{b_k[0]}{\sqrt{2}}, \quad b_2^k[0] = \frac{-b_k^*[1]}{\sqrt{2}}$$
$$b_1^k[1] = \frac{b_k[1]}{\sqrt{2}}, \quad b_2^k[1] = \frac{b_k^*[0]}{\sqrt{2}}$$

The additional normalization by $\sqrt{2}$ has been introduced so that the total transmitted energy remains constant and independent of M. In the receiver we employ two linear filters. The first, \mathbf{W}_1 , is trained to extract the evennumbered symbols $(b_1[0], b_1[2], \ldots)$ and the other, \mathbf{W}_2 , is trained to extract the odd-numbered symbols (in this case user number one is the desired user). **4.2.1.** Least-Squares Optimization. Using least squares, the solution then becomes

$$\mathbf{W}_{1}^{\text{LS}} = \arg\min_{\mathbf{W}} \frac{2}{N_{t}} \sum_{i=0}^{N_{t}/2-1} |\mathbf{W}^{H}\mathbf{r}(i) - b_{1}[2i]|^{2}$$
(16)

$$\mathbf{W}_{2}^{\text{LS}} = \arg\min_{\mathbf{W}} \frac{2}{N_{t}} \sum_{i=0}^{N_{t}/2-1} |\mathbf{W}^{H}\mathbf{r}(i) - b_{1}^{*}[2i+1]|^{2} \quad (17)$$

where N_t is the number of training symbols per burst (see Fig. 3) and

$$\mathbf{r}(i) = [\mathbf{r}_1(2i)^T, \mathbf{r}_1^*(2i+1)^T, \dots, r_A(2i)^T, r_A^*(2i+1)^T]^T$$

The second vector for each antenna element is conjugated to account for the conjugate introduced by the spacetime code.

If $N_t/2 \ge 2AN$, then the solution to Eqs. (16) and (17) is well known:

$$\mathbf{W}_{1}^{\text{LS}} = \underbrace{\left(\frac{2}{N_{t}}\sum_{i=0}^{N_{t}/2-1}\mathbf{r}(i)\mathbf{r}(i)^{H}\right)^{-1}}_{\mathbf{W}_{2}^{\text{LS}}} \underbrace{\left(\frac{2}{N_{t}}\sum_{i=0}^{N_{t}/2-1}\mathbf{r}(i)b_{1}^{*}[2i]\right)}_{\mathbf{W}_{2}^{\text{LS}}} = \underbrace{\left(\frac{2}{N_{t}}\sum_{i=0}^{N_{t}/2-1}\mathbf{r}(i)\mathbf{r}(i)^{H}\right)^{-1}}_{\mathbf{V}_{t}} \underbrace{\left(\frac{2}{N_{t}}\sum_{i=0}^{N_{t}/2-1}\mathbf{r}(i)b_{1}[2i+1]\right)}_{\mathbf{P}_{N_{t}^{2}}} (18)$$

In cases where this is not true, the time-averaged autocorrelation matrix, \mathbf{R}_{N_l} , becomes singular, so we use diagonal loading for regularization

$$\mathbf{W}_{1,2}^{\text{LS}} = (\mathbf{R}_{N_t} + \delta \mathbf{I}_{2AN})^{-1} \mathbf{P}_{N_t}^{1,2}$$
(20)

п

where δ is some small positive constant (σ^2 ideally). As $N_t \to \infty$, $\mathbf{R}_{N_t} \to \mathbf{R}$, where \mathbf{R} is the autocorrelation matrix

$$\mathbf{R} = \mathbf{G}\mathbf{B}\mathbf{G}^H + \sigma^2 \mathbf{I}_{2AN} \tag{21}$$



Figure 3. Burst structure.

and

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_1^* \\ \vdots & \vdots \\ \mathbf{G}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_A^* \end{bmatrix}, \quad \mathbf{B} = E\left(\begin{bmatrix} \mathbf{b}_{2i} \\ \mathbf{b}_{2i+1}^* \end{bmatrix} \begin{bmatrix} \mathbf{b}_{2i}^H & \mathbf{b}_{2i+1}^T \end{bmatrix}\right)$$

Assuming that $2AN \ge 4KL$, and that **G** has full rank, it can be shown that the dimensionality of the signal space, k, (rank(**GBG**^H)) is given by

$$\kappa = \begin{cases} 2 \cdot K & \text{if } L = 1 \text{ (flat-fading)} \\ 2 \cdot K \cdot (L+1) & \text{if } L > 1 \end{cases}$$
(22)

Thus, in the noiseless case ($\sigma^2 = 0$) the minimum number of training symbols required to completely suppress the interference (multiple access and intersymbol) is simply $2 \cdot \kappa$. However, in the presence of noise and in a time-varying propagation environment, longer training sequences are needed. For a given length of training, the performance degrades with decreasing SNR and is highly dependent on the time-varying channel characteristics. This has serious implications on both user detectability and system spectral efficiency, which is measured as a percentage of training data in a burst. To alleviate this problem semiblind algorithms that utilize blind signal estimation principles to effectively increase the equivalent length of the training sequence have been proposed [1,14,16]. The semiblind algorithm described next can be effectively incorporated in a spatiotemporal framework and provide significant performance gains as it will be demonstrated by means of computer simulations.

4.2.2. Semiblind Processing. The objective is to find the weight vectors that minimize the cost functions

$$\mathbf{W}_{1} = \arg\min_{\mathbf{W}} \frac{2}{N_{b}} \sum_{i=0}^{N_{b}/2-1} |\mathbf{W}^{H}\mathbf{r}(i) - b_{1}[2i]|^{2}$$
(23)

$$\mathbf{W}_{2} = \arg\min_{\mathbf{W}} \frac{2}{N_{b}} \sum_{i=0}^{N_{b}/2-1} |\mathbf{W}^{H}\mathbf{r}(i) - b_{1}^{*}[2i+1]|^{2} \quad (24)$$

given the following information about the source symbols:

*b*₁[*i*] is known for *i* = 0, ..., *N_t* - 1 (training symbols),
|*b*₁[*i*]| = 1, *i* = 0, ..., *N_b* - 1.

Our strategy is to use an alternating projection approach such as that discussed by van der Veen [2]. For the *i*th iteration of the weight vector, let

$$\tilde{b}_{1}^{1(i)} \triangleq \left\{ \mathbf{b}_{N_{t}}^{1}, \frac{\mathbf{W}_{1}^{H(i)}\mathbf{r}(N_{t}/2)}{|\mathbf{W}_{1}^{H(i)}\mathbf{r}(N_{t}/2)|}, \dots, \frac{\mathbf{W}_{1}^{H(i)}\mathbf{r}(N_{b}/2-1)}{|\mathbf{W}_{1}^{H(i)}\mathbf{r}(N_{b}/2-1)|} \right\}$$
$$\tilde{b}_{1}^{2(i)} \triangleq \left\{ \mathbf{b}_{N_{t}}^{2}, \frac{\mathbf{W}_{2}^{H(i)}\mathbf{r}(N_{t}/2)}{|\mathbf{W}_{2}^{H(i)}\mathbf{r}(N_{t}/2)|}, \dots, \frac{\mathbf{W}_{2}^{H(i)}\mathbf{r}(N_{b}/2-1)}{|\mathbf{W}_{2}^{H(i)}\mathbf{r}(N_{b}/2-1)|} \right\}$$

where $\mathbf{b}_{N_t}^1 = [b_1[0], b_1[2], \dots, b_1[N_t - 2]]$ and $\mathbf{b}_{N_t}^2 = [b_1[1], b_1[3], \dots, b_1[N_t - 1]])$ are a sequence that contains

the known N_t source symbols, and $(N_b - N_t)/2$ estimated source symbols (via the Godard-type nonlinearity $\frac{\mathbf{W}_1^H \mathbf{r}(k)}{|\mathbf{W}_1^H \mathbf{r}(k)|}$). The function $\frac{\mathbf{W}_1^H \mathbf{r}(k)}{|\mathbf{W}_1^H \mathbf{r}(k)|}$ is chosen based on a priori knowledge of the source constellation (it returns a complex number with unit magnitude). If the initialization is accurate, then it will return the correct phase of the source symbol. We refer to the corresponding $(N_b - N_t)/2$ symbols in $\tilde{b}_1^{1(t)}(\tilde{b}_1^{2(t)})$ as pseudotraining symbols.

symbols in $\tilde{b}_{1}^{1(i)}(\tilde{b}_{1}^{2(i)})$ as *pseudotraining symbols*. The sequence $\tilde{b}_{1}^{1(i)}(\tilde{b}_{1}^{2(i)})$ is then used to estimate a new cross-correlation vector, $\tilde{\mathbf{P}}_{N_b}^{1(i)}(\tilde{\mathbf{P}}_{N_b}^{2(i)})$, and $\mathbf{W}_{1}^{(i)}(\mathbf{W}_{2}^{(i)})$ is updated using this new vector and signal space parameters computed from eigen-decomposition of the time-averaged autocorrelation matrix \mathbf{R}_{N_b} :

$$\tilde{\mathbf{P}}_{N_b}^{1(i)} = \frac{2}{N_b} \sum_{k=0}^{N_b/2-1} \tilde{b}_1^{*1(i)}(k) \mathbf{r}(k)$$
(25)

$$\tilde{\mathbf{P}}_{N_b}^{2(i)} = \frac{2}{N_b} \sum_{k=0}^{N_b/2-1} \tilde{b}_1^{2(i)}(k) \mathbf{r}(k)$$
(26)

$$\mathbf{W}_{1}^{(i+1)} = \hat{\mathbf{U}}_{S} \hat{\mathbf{\Lambda}}_{S}^{-1} \hat{\mathbf{U}}_{S}^{H} \tilde{\mathbf{P}}_{N_{b}}^{1(i)}$$

$$\tag{27}$$

$$\mathbf{W}_{2}^{(i+1)} = \hat{\mathbf{U}}_{S} \hat{\mathbf{\Lambda}}_{S}^{-1} \hat{\mathbf{U}}_{S}^{H} \tilde{\mathbf{P}}_{N_{b}}^{2(i)}$$
(28)

where \mathbf{U}_S are the signal space eigenvectors and Λ_S are the signal space eigenvalues. The complete algorithm is given in Table 1. The algorithm is called the (*semiblind constant-modulus algorithm*) (SBCMA).

5. SIMULATION EXAMPLES

This section compares the MSE and probability of error performance of the SBCMA algorithm against LS. Channel coefficients, $\{\beta_{ia}^k[l]\}_{l=0}^{L_k-1}$, are randomly generated for each user $(k = 1, \ldots, K)$ and for each transmitter/receiver pair $(i = 1, 2, a = 1, \ldots, A)$ from a complex Gaussian distribution of unit variance and zero mean. The channel coefficients do not change during the transmission of N_b data symbols. SNR is defined as $10 \log_{10}(1/\sigma^2)$. In all cases

$$\begin{split} & In: \hat{\mathbf{U}}_{s}, \hat{\mathbf{A}}_{s}, \{\mathbf{r}(k)\}_{k=0}^{N_{b}/2-1}, \{b_{1}[k]\}_{k=0}^{N_{t}-1}, \zeta \quad Out: \mathbf{W}_{1,2} \\ & \mathbf{H} \triangleq \hat{\mathbf{U}}_{s} \hat{\mathbf{A}}_{s}^{-1} \hat{\mathbf{U}}_{s}^{H} \\ & \mathbf{X} \triangleq [\mathbf{r}(0), \dots, \mathbf{r}(N_{b}/2-1)] \\ & \text{Choose } \mathbf{W}_{1}^{(0)} = \mathbf{W}_{1}^{LS}, \mathbf{W}_{2}^{(0)} = \mathbf{W}_{2}^{LS} \\ & \text{for } i = 0, 1, \dots \\ & \text{a. } Y_{1} = (\mathbf{W}_{1}^{(i)H}\mathbf{X})/|\mathbf{W}_{1}^{(i)H}\mathbf{X}| \\ & Y_{2} = (\mathbf{W}_{2}^{(i)H}\mathbf{X})/|\mathbf{W}_{2}^{(i)H}\mathbf{X}| \\ & \tilde{\mathbf{b}}_{1} = [\mathbf{b}_{N_{t}}^{1}, Y_{1}[N_{t}/2], \dots, Y_{1}[N_{b}/2-1]] \\ & \tilde{\mathbf{b}}_{2} = [\mathbf{b}_{N_{t}}^{2}, Y_{2}[N_{t}/2], \dots, Y_{2}[N_{b}/2-1]] \\ & \tilde{\mathbf{b}}. \quad \tilde{\mathbf{P}}_{N_{b}}^{1(i)} = (\mathbf{X} \cdot \tilde{\mathbf{b}}_{1}^{H})/(N_{b}/2) \\ & \tilde{\mathbf{P}}_{N_{b}}^{2(i)} = (\mathbf{X} \cdot \tilde{\mathbf{b}}_{2}^{T})/(N_{b}/2) \\ & \text{c. } \mathbf{W}_{1}^{(i+1)} = \mathbf{H} \cdot \tilde{\mathbf{P}}_{N_{b}}^{1(i)}, \mathbf{W}_{2}^{(i+1)} = \mathbf{H} \cdot \tilde{\mathbf{P}}_{N_{b}}^{2(i)} \\ & \text{until } \|\mathbf{W}_{1}^{(i+1)} - \mathbf{W}_{1}^{(i)}\|^{2}/\|\mathbf{W}_{1}^{(i)}\|^{2} < \zeta \end{split}$$



Figure 4. MSE comparison for noiseless case ($\sigma^2 = 0$): L = 2; K = 1, 2, 3, 4; A = 2; N = 7.

 $\zeta = 10^{-5}$, and N = 7. The signal space rank was estimated using the MDL criterion.

Figure 4 demonstrates the zero-forcing ability of the LS filter when the number of training symbols is greater than the signal space dimensionality [as predicted by Eq. (22)]. In this case L = 2 and K is increased from 1 to 4. It is seen that once $N_t \geq 2 \cdot \kappa$ there is complete suppression of multiple access and intersymbol interference.

Figure 5 compares the MSE performance of the SBCMA with LS under the conditions that K = 3, L = 2 and A = 2 or A = 4. It is observed that the SBCMA uses significantly fewer training symbols than the LS and that the SBCMA is able to converge using fewer training symbols than required by LS in the noiseless case (36 in this case). When A = 4, the MSE is significantly lower but the required number of training symbols for convergence is roughly the



Figure 5. MSE comparison for A = 2 and A = 4: SNR = 10 dB, L = 2, K = 3, N = 7.



Figure 6. Probability of error comparison for A = 2 and A = 4: SNR = 10 dB, L = 2, K = 3, N = 7.

same. This can be understood from the fact that increasing A does not increase the column space of **G**.

Finally, Fig. 6 compares the probability of error performance of the SBCMA with LS under the same conditions as above. Again, we observe the superior convergence speed of the SBCMA over LS.

6. CONCLUSION

Some theoretical results and a practical approach on spatio-temporal signal processing for wireless communications have been presented. The investigation for efficient utilization of multiple antennas at both ends of a wireless transceiver combined has raised the expectations for significant increases of the system capacity but at the same time created new challenges in developing the required signal processing technologies. The research and developments in this area are ongoing, and the corresponding literature is rapidly increasing as we now try to define the systems that will succeed the third generation of wireless mobile systems.

BIOGRAPHIES

Dimitrios Hatzinakos, Ph.D., is a Professor at the Department of Electrical and Computer Engineering, University of Toronto. He has also served as Chair of the Communications Group of the Department since July 1, 1999. His research interests are in the area of digital signal processing with applications to wireless communications, image processing, and multimedia. He is author or co-author of more than 120 papers in technical journals and conference proceedings, and he has contributed to five books in his areas of interest.

He served as an Associate Editor for the *IEEE Transactions on Signal Processing* from July 1998 until 2001; Guest Editor for the special issue of *Signal Processing*, Elsevier, on *Signal Processing Technologies*

2340 SPEECH CODING: FUNDAMENTALS AND APPLICATIONS

for Short Burst Wireless Communications, which appeared in October 2000. He was a member of the Conference board of the IEEE Statistical Signal and Array Processing Technical Committee (SSAP) from 1992 to 1995 and was Technical Program co-Chair of the 5th Workshop on Higher-Order Statistics in July 1997. He is a senior member of the IEEE and member of EURASIP, the Professional Engineers of Ontario (PEO), and the Technical Chamber of Greece.

Ryan A. Pacheco is currently working towards a Ph.D. in Electrical Engineering at the University of Toronto. His research is concerned with blind/semiblind interference suppression, tracking, and signal acquisition for wireless communications.

BIBLIOGRAPHY

- 1. R. A. Pacheco and D. Hatzinakos, Semi-blind strategies for interference suppression in ds-cdma systems, *IEEE Trans. Signal Process.* (in press).
- A.-J. van der Veen, Algebraic constant modulus algorithms, in Signal Processing Advances in Wireless and Mobile Communications, Vol. 2, Prentice-Hall PTR, Englewood Cliffs, NJ, 2001, pp. 89–130.
- G. Ganesan and P. Stoica, Space-time diversity, in G. B. Giannakis et al., eds., Signal Processing Advances in Wireless & Mobile Communications, Vol. 2, Prentice-Hall PTR, Englewood Cliffs, NJ, 2001, pp. 59–87.
- V. Tarokh, H. Jafarkhani, and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* 45(5): 1456-1467 (July 1999).
- A. F. Naguib, N. Seshadri, and A. R. Calderbank, Increasing data rate over wireless channels, *IEEE Signal Process. Mag.* 76–92 (May 2000).
- 6. S. Haykin, Adaptive Filter Theory Prentice-Hall, Englewood Cliffs, NJ, 1996.
- S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* 16(8): 1451-1458 (Oct. 1998).
- G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multiple antennas, *Bell Labs Tech. J.* 1(2): 41–59 (1996).
- P. W. Wolniansky, G. J. Foschini, G. B. Golden, and R. A. Valenzuela, V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel, *IEEE Proc. ISSSE* 295-300 (1998).
- M. Adachi, F. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communication systems, *IEEE Commun. Mag.* 36: 56-69 (Sept. 1998).
- A. Gorokhov and P. Loubaton, Semi-blind second order identification of convolutive channels, *Proc. ICASSP*, 1997, pp. 3905–3908.
- D. Hatzinakos, Blind deconvolution channel identification and equalization, *Control Dynam. Syst.* 68: 279-331 (1995).
- M. Honig, U. Madhow, and S. Verdu, Blind multiuser detection, *IEEE Trans. Inform. Theory* **41**: 944–960 (July 1995).
- 14. A. M. Kuzminskiy, L. Féty, P. Forster, and S. Mayrargue, Regularized semi-blind estimation of spatio-temporal filter

coefficients for mobile radio communications, *Proc. GRETSI*, 1997, pp. 127–130.

- A. M. Kuzminskiy and D. Hatzinakos, Semi-blind estimation of spatio-temporal filter coefficients based on a training-like approach, *IEEE Signal Process. Lett.* 5: 231–233 (Sept. 1998).
- A. M. Kuzminskiy and D. Hatzinakos, Multistage semi-blind spatio-temporal processing for short burst multiuser SDMA systems, Proc. 32nd Asilomar Conf. Signals, Systems, and Computers, Oct. 1998, pp. 1887–1891.
- A. Paulraj and C. Papadias, Space-time processing for wireless communications, *IEEE Signal Process. Mag.* 14: 49-83 (Nov. 1997).
- J. G. Proakis, *Digital Communications*, 3rd. ed., McGraw-Hill, New York, 1995.
- A. van der Veen, S. Talwar, and A. Paulraj, A subspace approach to blind space-time signal processing for wireless communication systems, *IEEE Trans. Signal Process.* 45: 173-190 (Jan. 1997).
- S. Verdu, Minimum probability of error for asynchronous gaussian multiple-access channels, *IEEE Trans. Inform. Theory* 32: 85–96 (Jan. 1986).
- X. Wang and H. V. Poor, Blind equalization and multiuser detection in dispersive CDMA channels, *IEEE Trans. Commun.* 46: 91-103 (Jan. 1998).

SPEECH CODING: FUNDAMENTALS AND APPLICATIONS

MARK HASEGAWA-JOHNSON University of Illinois at Urbana–Champaign Urbana, Illinois ABEER ALWAN University of California at Los Angeles Los Angeles, California

1. INTRODUCTION

Speech coding is the process of obtaining a compact representation of voice signals for efficient transmission over band-limited wired and wireless channels and/or storage. Today, speech coders have become essential components in telecommunications and in the multimedia infrastructure. Commercial systems that rely on efficient speech coding include cellular communication, voice over internet protocol (VOIP), videoconferencing, electronic toys, archiving, and digital simultaneous voice and data (DSVD), as well as numerous PC-based games and multimedia applications.

Speech coding is the art of creating a minimally redundant representation of the speech signal that can be efficiently transmitted or stored in digital media, and decoding the signal with the best possible perceptual quality. Like any other continuous-time signal, speech may be represented digitally through the processes of sampling and quantization; speech is typically quantized using either 16-bit uniform or 8-bit companded quantization. Like many other signals, however, a sampled speech signal contains a great deal of information that is either redundant (nonzero mutual information between successive samples in the signal) or perceptually irrelevant (information that is not perceived by human listeners). Most telecommunications coders are *lossy*, meaning that the synthesized speech is perceptually similar to the original but may be physically dissimilar.

A speech coder converts a digitized speech signal into a coded representation, which is usually transmitted in frames. A speech decoder receives coded frames and synthesizes reconstructed speech. Standards typically dictate the input-output relationships of both coder and decoder. The input-output relationship is specified using a reference implementation, but novel implementations are allowed, provided that input-output equivalence is maintained. Speech coders differ primarily in bit rate (measured in bits per sample or bits per second), complexity (measured in operations per second), delay (measured in milliseconds between recording and playback), and perceptual quality of the synthesized speech. Narrowband (NB) coding refers to coding of speech signals whose bandwidth is less than 4 kHz (8 kHz sampling rate), while wideband (WB) coding refers to coding of 7-kHz-bandwidth signals (14-16 kHz sampling rate). NB coding is more common than WB coding mainly because of the narrowband nature of the wireline telephone channel (300-3600 Hz). More recently, however, there has been an increased effort in wideband speech coding because of several applications such as videoconferencing.

There are different types of speech coders. Table 1 summarizes the bit rates, algorithmic complexity, and standardized applications of the four general classes of coders described in this article; Table 2 lists a selection of specific speech coding standards. Waveform coders attempt to code the exact shape of the speech signal waveform, without considering the nature of human speech production and speech perception. These coders are high-bit-rate coders (typically above 16 kbps). Linear prediction coders (LPCs), on the other hand, assume that the speech signal is the output of a linear time-invariant (LTI) model of speech production. The transfer function of that model is assumed to be all-pole (autoregressive model). The excitation function is a quasiperiodic signal constructed from discrete pulses (1-8 per pitch period), pseudorandom noise, or some combination of the two. If the excitation is generated only at the receiver, based on a transmitted pitch period and voicing information, then the system is designated as an LPC vocoder. LPC vocoders that provide extra information about the spectral shape of the excitation have been adopted as coder standards between 2.0 and 4.8 kbps. LPC-based analysis-by-synthesis coders

(LPC-AS), on the other hand, choose an excitation function by explicitly testing a large set of candidate excitations and choosing the best. LPC-AS coders are used in most standards between 4.8 and 16 kbps. Subband coders are frequency-domain coders that attempt to parameterize the speech signal in terms of spectral properties in different frequency bands. These coders are less widely used than LPC-based coders but have the advantage of being scalable and do not model the incoming signal as speech. Subband coders are widely used for high-quality audio coding.

This article is organized as follows. Sections 2, 3, 4 and 5 present the basic principles behind waveform coders, subband coders, LPC-based analysis-by-synthesis coders, and LPC-based vocoders, respectively. Section 6 describes the different quality metrics that are used to evaluate speech coders, while Section 7 discusses a variety of issues that arise when a coder is implemented in a communications network, including voice over IP, multirate coding, and channel coding. Section 8 presents an overview of standardization activities involving speech coding, and we conclude in Section 9 with some final remarks.

2. WAVEFORM CODING

Waveform coders attempt to code the exact shape of the speech signal waveform, without considering in detail the nature of human speech production and speech perception. Waveform coders are most useful in applications that require the successful coding of both speech and nonspeech signals. In the public switched telephone network (PSTN), for example, successful transmission of modem and fax signaling tones, and switching signals is nearly as important as the successful transmission of speech. The most commonly used waveform coding algorithms are uniform 16-bit PCM, companded 8-bit PCM [48], and ADPCM [46].

2.1. Pulse Code Modulation (PCM)

Pulse code modulation (PCM) is the name given to memoryless coding algorithms that quantize each sample of s(n) using the same reconstruction levels \hat{s}_k , $k = 0, \ldots, m, \ldots, K$, regardless of the values of previous samples. The reconstructed signal $\hat{s}(n)$ is given by

$$\hat{s}(n) = \hat{s}_m$$
 for $s(n)$ s.t. $(s(n) - \hat{s}_m)^2 = \min_{k=0,\dots,K} (s(n) - \hat{s}_k)^2$
(1)

Many speech and audio applications use an odd number of reconstruction levels, so that background noise signals with a very low level can be quantized exactly to

 Table 1. Characteristics of Standardized Speech Coding Algorithms in Each of

 Four Broad Categories

Speech Coder Class	Rates (kbps)	Complexity	Standardized Applications	Section
Waveform coders	16 - 64	Low	Landline telephone	2
Subband coders	12 - 256	Medium	Teleconferencing, audio	3
LPC-AS	4.8 - 16	High	Digital cellular	4
LPC vocoder	2.0 - 4.8	High	Satellite telephony, military	5

	Rate	\mathbf{BW}	Standards	Standard		
Application	(kbps)	(kHz)	Organization	Number	Algorithm	Year
Landline telephone	64	3.4	ITU	G.711	μ -law or A-law PCM	1988
	16 - 40	3.4	ITU	G.726	ADPCM	1990
	16 - 40	3.4	ITU	G.727	ADPCM	1990
Tele conferencing	48 - 64	7	ITU	G.722	Split-band ADPCM	1988
	16	3.4	ITU	G.728	Low-delay CELP	1992
Digital	13	3.4	ETSI	Full-rate	RPE-LTP	1992
cellular	12.2	3.4	ETSI	\mathbf{EFR}	ACELP	1997
	7.9	3.4	TIA	IS-54	VSELP	1990
	6.5	3.4	ETSI	Half-rate	VSELP	1995
	8.0	3.4	ITU	G.729	ACELP	1996
	4.75 - 12.2	3.4	ETSI	AMR	ACELP	1998
	1 - 8	3.4	CDMA-TIA	IS-96	QCELP	1993
Multimedia	5.3 - 6.3	3.4	ITU	G.723.1	MPLPC, CELP	1996
	2.0 - 18.2	3.4 - 7.5	ISO	MPEG-4	HVXC, CELP	1998
Satellite telephony	4.15	3.4	INMARSAT	Μ	IMBE	1991
	3.6	3.4	INMARSAT	Mini-M	AMBE	1995
Secure communications	2.4	3.4	DDVPC	FS1015	LPC-10e	1984
	2.4	3.4	DDVPC	MELP	MELP	1996
	4.8	3.4	DDVPC	FS1016	CELP	1989
	16 - 32	3.4	DDVPC	CVSD	CVSD	

Table 2. A Representative Sample of Speech Coding Standards

 $\hat{s}_{K/2} = 0$. One important exception is the A-law companded PCM standard [48], which uses an even number of reconstruction levels.

2.1.1. Uniform PCM. Uniform PCM is the name given to quantization algorithms in which the reconstruction levels are uniformly distributed between $S_{\rm max}$ and $S_{\rm min}$. The advantage of uniform PCM is that the quantization error power is independent of signal power; high-power signals are quantized with the same resolution as low-power signals. Invariant error power is considered desirable in many digital audio applications, so 16-bit uniform PCM is a standard coding scheme in digital audio.

The error power and SNR of a uniform PCM coder vary with bit rate in a simple fashion. Suppose that a signal is quantized using *B* bits per sample. If zero is a reconstruction level, then the quantization step size Δ is

$$\Delta = \frac{S_{\max} - S_{\min}}{2^B - 1} \tag{2}$$

Assuming that quantization errors are uniformly distributed between $\Delta/2$ and $-\Delta/2$, the quantization error power is

$$10 \log_{10} E[e^2(n)] = 10 \log_{10} \frac{\Delta^2}{12}$$

\$\approx constant + 20 \log_{10}(S_{max} - S_{min}) - 6B\$ (3)

2.1.2. Companded PCM. Companded PCM is the name given to coders in which the reconstruction levels \hat{s}_k are not uniformly distributed. Such coders may be modeled using a compressive nonlinearity, followed by uniform PCM, followed by an expansive nonlinearity:

$$s(n) \rightarrow$$
 compress $\rightarrow t(n) \rightarrow$ uniform PCM
 $\rightarrow \hat{t}(n) \rightarrow$ expand $\rightarrow \hat{s}(n)$ (4)

It can be shown that, if small values of s(n) are more likely than large values, expected error power is minimized by a companding function that results in a higher density of reconstruction levels \hat{x}_k at low signal levels than at high signal levels [78]. A typical example is the μ -law companding function [48] (Fig. 1), which is given by

$$t(n) = S_{\max} \frac{\log(1+\mu|s(n)/S_{\max}|)}{\log(1+\mu)} \operatorname{sign}(s(n))$$
(5)

where μ is typically between 0 and 256 and determines the amount of nonlinear compression applied.

2.2. Differential PCM (DPCM)

Successive speech samples are highly correlated. The longterm average spectrum of voiced speech is reasonably



Figure 1. μ -law companding function, $\mu = 0, 1, 2, 4, 8, \dots, 256$.



Figure 2. Schematic of a DPCM coder.

well approximated by the function S(f) = 1/f above about 500 Hz; the first-order intersample correlation coefficient is approximately 0.9. In differential PCM, each sample s(n) is compared to a prediction $s_p(n)$, and the difference is called the prediction residual d(n) (Fig. 2). d(n) has a smaller dynamic range than s(n), so for a given error power, fewer bits are required to quantize d(n).

Accurate quantization of d(n) is useless unless it leads to accurate quantization of s(n). In order to avoid amplifying the error, DPCM coders use a technique copied by many later speech coders; the encoder includes an embedded decoder, so that the reconstructed signal $\hat{s}(n)$ is known at the encoder. By using $\hat{s}(n)$ to create $s_p(n)$, DPCM coders avoid amplifying the quantization error:

$$d(n) = s(n) - s_p(n) \tag{6}$$

$$\hat{s}(n) = \hat{d}(n) + s_p(n) \tag{7}$$

$$e(n) = s(n) - \hat{s}(n) = d(n) - \hat{d}(n)$$
(8)

Two existing standards are based on DPCM. In the first type of coder, continuously varying slope delta modulation (CVSD), the input speech signal is upsampled to either 16 or 32 kHz. Values of the upsampled signal are predicted using a one-tap predictor, and the difference signal is quantized at one bit per sample, with an adaptively varying Δ . CVSD performs badly in quiet environments, but in extremely noisy environments (e.g., helicopter cockpit), CVSD performs better than any LPC-based algorithm, and for this reason it remains the U.S. Department of Defense recommendation for extremely noisy environments [64,96].

DPCM systems with adaptive prediction and quantization are referred to as adaptive differential PCM systems (ADPCM). A commonly used ADPCM standard is G.726, which can operate at 16, 24, 32, or 40 kbps (2-5 bits/sample) [45]. G.726 ADPCM is frequently used at 32 kbps in landline telephony. The predictor in G.726 consists of an adaptive second-order IIR predictor in series with an adaptive sixth-order FIR predictor. Filter coefficients are adapted using a computationally simplified gradient descent algorithm. The prediction residual is quantized using a semilogarithmic companded PCM quantizer at a rate of 2-5 bits per sample. The quantization step size adapts to the amplitude of previous samples of the quantized prediction error signal; the speed of adaptation is controlled by an estimate of the type of signal, with adaptation to speech signals being faster than adaptation to signaling tones.

3. SUBBAND CODING

In subband coding, an analysis filterbank is first used to filter the signal into a number of frequency bands and then bits are allocated to each band by a certain criterion. Because of the difficulty in obtaining high-quality speech at low bit rates using subband coding schemes, these techniques have been used mostly for wideband medium to high bit rate speech coders and for audio coding.

For example, G.722 is a standard in which ADPCM speech coding occurs within two subbands, and bit allocation is set to achieve 7-kHz audio coding at rates of 64 kbps or less.

In Refs. 12,13, and 30 subband coding is proposed as a flexible scheme for robust speech coding. A speech production model is not used, ensuring robustness to speech in the presence of background noise, and to nonspeech sources. High-quality compression can be achieved by incorporating masking properties of the human auditory system [54,93]. In particular, Tang et al. [93] present a scheme for robust, high-quality, scalable, and embedded speech coding. Figure 3 illustrates the basic structure of the coder. Dynamic bit allocation and prioritization and embedded quantization are used to optimize the perceptual quality of the embedded bitstream, resulting in little performance degradation relative to a nonembedded implementation. A subband spectral analysis technique was developed that substantially reduces the complexity of computing the perceptual model.

The encoded bitstream is embedded, allowing the coder output to be scalable from high quality at higher



Figure 3. Structure of a perceptual subband speech coder [93].

2344 SPEECH CODING: FUNDAMENTALS AND APPLICATIONS

bit rates, to lower quality at lower rates, supporting a wide range of service and resource utilization. The lower bit rate representation is obtained simply through truncation of the higher bit rate representation. Since source rate adaptation is performed through truncation of the encoded stream, interaction with the source coder is not required, making the coder ideally suited for rate adaptive communication systems.

Even though subband coding is not widely used for speech coding today, it is expected that new standards for wideband coding and rate-adaptive schemes will be based on subband coding or a hybrid technique that includes subband coding. This is because subband coders are more easily scalable in bit rate than standard CELP techniques, an issue which will become more critical for high-quality speech and audio transmission over wireless communication channels and the Internet, allowing the system to seamlessly adapt to changes in both the transmission environment and network congestion.

4. LPC-BASED ANALYSIS BY SYNTHESIS

An analysis-by-synthesis speech coder consists of the following components:

• A model of speech production that depends on certain parameters θ :

$$\hat{s}(n) = f(\theta) \tag{9}$$

• A list of *K* possible parameter sets for the model

$$\theta_1,\ldots,\theta_k,\ldots,\theta_K$$
 (10)

• An error metric $|E_k|^2$ that compares the original speech signal s(n) and the coded speech signal $\hat{s}(n)$. In LPC-AS coders, $|E_k|^2$ is typically a perceptually weighted mean-squared error measure.

A general analysis-by-synthesis coder finds the optimum set of parameters by synthesizing all of the Kdifferent speech waveforms $\hat{s}_k(n)$ corresponding to the K possible parameter sets θ_k , computing $|E_k|^2$ for each synthesized waveform, and then transmitting the index of the parameter set which minimizes $|E_k|^2$. Choosing a set of transmitted parameters by explicitly computing $\hat{s}_k(n)$ is called "closed loop" optimization, and may be contrasted with "open-loop" optimization, in which coder parameters are chosen on the basis of an analytical formula without explicit computation of $\hat{s}_k(n)$. Closed-loop optimization of all parameters is prohibitively expensive, so LPC-based analysis-by-synthesis coders typically adopt the following compromise. The gross spectral shape is modeled using an all-pole filter 1/A(z) whose parameters are estimated in open-loop fashion, while spectral fine structure is modeled using an excitation function U(z) whose parameters are optimized in closed-loop fashion (Fig. 4).

4.1. The Basic LPC Model

In LPC-based coders, the speech signal S(z) is viewed as the output of a linear time-invariant (LTI) system whose



Figure 4. General structure of an LPC-AS coder (**a**) and decoder (**b**). LPC filter A(z) and perceptual weighting filter W(z) are chosen open-loop, then the excitation vector u(n) is chosen in a closed-loop fashion in order to minimize the error metric $|E|^2$.

input is the excitation signal U(z), and whose transfer function is represented by the following:

$$S(z) = \frac{U(z)}{A(z)} = \frac{U(z)}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$
(11)

Most of the zeros of A(z) correspond to resonant frequencies of the vocal tract or *formant frequencies*. Formant frequencies depend on the geometry of the vocal tract; this is why men and women, who have different vocal-tract shapes and lengths, have different formant frequencies for the same sounds.

The number of LPC coefficients (p) depends on the signal bandwidth. Since each pair of complex-conjugate poles represents one formant frequency and since there is, on average, one formant frequency per 1 kHz, p is typically equal to 2BW (in kHz) + (2 to 4). Thus, for a 4 kHz speech signal, a 10th-12th-order LPC model would be used.

This system is excited by a signal u(n) that is uncorrelated with itself over lags of less than p + 1. If the underlying speech sound is unvoiced (the vocal folds do not vibrate), then u(n) is uncorrelated with itself even at larger time lags, and may be modeled using a pseudorandom-noise signal. If the underlying speech is voiced (the vocal folds vibrate), then u(n) is quasiperiodic with a fundamental period called the "pitch period."

4.2. Pitch Prediction Filtering

In an LPC-AS coder, the LPC excitation is allowed to vary smoothly between fully voiced conditions (as in a vowel) and fully unvoiced conditions (as in /s/). Intermediate levels of voicing are often useful to model partially voiced phonemes such as /z/.

The partially voiced excitation in an LPC-AS coder is constructed by passing an uncorrelated noise signal c(n)through a pitch prediction filter [2,79]. A typical pitch prediction filter is

$$u(n) = gc(n) + bu(n - T_0)$$
(12)

where T_0 is the pitch period. If c(n) is unit variance white noise, then according to Eq. (12) the spectrum of u(n) is

$$|U(e^{j\omega})|^2 = \frac{g^2}{1 + b^2 - 2b\cos\omega T_0}$$
(13)

Figure 5 shows the normalized magnitude spectrum $(1-b)|U(e^{j\omega})|$ for several values of *b* between 0.25 and 1. As shown, the spectrum varies smoothly from a uniform spectrum, which is heard as unvoiced, to a harmonic spectrum that is heard as voiced, without the need for a binary voiced/unvoiced decision.

In LPC-AS coders, the noise signal c(n) is chosen from a "stochastic codebook" of candidate noise signals. The stochastic codebook index, the pitch period, and the gains b and g are chosen in a closed-loop fashion in order to minimize a perceptually weighted error metric. The search for an optimum T_0 typically uses the same algorithm as the search for an optimum c(n). For this reason, the list of excitation samples delayed by different candidate values of T_0 is typically called an "adaptive codebook" [87].

4.3. Perceptual Error Weighting

Not all types of distortion are equally audible. Many types of speech coders, including LPC-AS coders, use simple models of human perception in order to minimize the audibility of different types of distortion. In LPC-AS coding, two types of perceptual weighting are commonly used. The first type, perceptual weighting of the residual quantization error, is used during the LPC excitation search in order to choose the excitation vector with the least audible quantization error. The second type, adaptive postfiltering, is used to reduce the perceptual importance of any remaining quantization error.

4.3.1. Perceptual Weighting of the Residual Quantization Error. The excitation in an LPC-AS coder is chosen to minimize a perceptually weighted error metric. Usually,



Figure 5. Normalized magnitude spectrum of the pitch prediction filter for several values of the prediction coefficient.

the error metric is a function of the time domain waveform error signal

$$e(n) = s(n) - \hat{s}(n) \tag{14}$$

Early LPC-AS coders minimized the mean-squared error

$$\sum_{n} e^{2}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^{2} d\omega$$
 (15)

It turns out that the MSE is minimized if the error spectrum, $E(e^{j\omega})$, is white—that is, if the error signal e(n) is an uncorrelated random noise signal, as shown in Fig. 6.

Not all noises are equally audible. In particular, noise components near peaks of the speech spectrum are hidden by a "masking spectrum" $M(e^{i\omega})$, so that a shaped noise spectrum at lower SNR may be less audible than a white-noise spectrum at higher SNR (Fig. 7). The audibility of noise may be estimated using a noise-to-masker ratio $|E_w|^2$:

$$|E_w|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|E(e^{j\omega})|^2}{|M(e^{j\omega})|^2} d\omega$$
(16)

The masking spectrum $M(e^{i\omega})$ has peaks and valleys at the same frequencies as the speech spectrum, but the difference in amplitude between peaks and valleys is somewhat smaller than that of the speech spectrum. A variety of algorithms exist for estimating the masking spectrum, ranging from extremely simple to extremely complex [51]. One of the simplest model masking spectra that has the properties just described is as follows [2]:

$$M(z) = \frac{|A(z/\gamma_2)|}{|A(z/\gamma_1)|}, \quad 0 < \gamma_2 < \gamma_1 \le 1$$
(17)

where 1/A(z) is an LPC model of the speech spectrum. The poles and zeros of M(z) are at the same frequencies as the poles of 1/A(z), but have broader bandwidths. Since the zeros of M(z) have broader bandwidth than its poles, M(z) has peaks where 1/A(z) has peaks, but the difference between peak and valley amplitudes is somewhat reduced.



Figure 6. The minimum-energy quantization noise is usually characterized as white noise.



Figure 7. Shaped quantization noise may be less audible than white quantization noise, even at slightly lower SNR.

The noise-to-masker ratio may be efficiently computed by filtering the speech signal using a perceptual weighting filter W(z) = 1/M(z). The perceptually weighted input speech signal is

$$S_w(z) = W(z)S(z) \tag{18}$$

Likewise, for any particular candidate excitation signal, the perceptually weighted output speech signal is

$$\hat{S}_w(z) = W(z)\hat{S}(z) \tag{19}$$

Given $s_w(n)$ and $\hat{s}_w(n)$, the noise-to-masker ratio may be computed as follows:

$$|E_w|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S_w(e^{j\omega}) - \hat{S}_w(e^{j\omega})|^2 d\omega = \sum_n (s_w(n) - \hat{s}_w(n))^2$$
(20)

4.3.2. Adaptive Postfiltering. Despite the use of perceptually weighted error minimization, the synthesized speech coming from an LPC-AS coder may contain audible quantization noise. In order to minimize the perceptual effects of this noise, the last step in the decoding process is often a set of adaptive postfilters [11,80]. Adaptive postfiltering improves the perceptual quality of noisy speech by giving a small extra emphasis to features of the spectrum that are important for human-to-human communication, including the pitch periodicity (if any) and the peaks in the spectral envelope.

A pitch postfilter (or long-term predictive postfilter) enhances the periodicity of voiced speech by applying either an FIR or IIR comb filter to the output. The time delay and gain of the comb filter may be set equal to the transmitted pitch lag and gain, or they may be recalculated at the decoder using the reconstructed signal $\hat{s}(n)$. The pitch postfilter is applied only if the proposed comb filter gain is above a threshold; if the comb filter gain is below threshold, the speech is considered unvoiced, and no pitch postfilter is used. For improved perceptual quality, the LPC excitation signal may be interpolated to a higher sampling rate in order to allow the use of fractional pitch periods; for example, the postfilter in the ITU G.729 coder uses pitch periods quantized to $\frac{1}{8}$ sample. A short-term predictive postfilter enhances peaks in the spectral envelope. The form of the short-term postfilter is similar to that of the masking function M(z) introduced in the previous section; the filter has peaks at the same frequencies as 1/A(z), but the peak-to-valley ratio is less than that of A(z).

Postfiltering may change the gain and the average spectral tilt of $\hat{s}(n)$. In order to correct these problems, systems that employ postfiltering may pass the final signal through a one-tap FIR preemphasis filter, and then modify its gain, prior to sending the reconstructed signal to a D/A converter.

4.4. Frame-Based Analysis

The characteristics of the LPC excitation signal u(n)change quite rapidly. The energy of the signal may change from zero to nearly full amplitude within one millisecond at the release of a plosive sound, and a mistake of more than about 5 ms in the placement of such a sound is clearly audible. The LPC coefficients, on the other hand, change relatively slowly. In order to take advantage of the slow rate of change of LPC coefficients without sacrificing the quality of the coded residual, most LPC-AS coders encode speech using a frame-subframe structure, as depicted in Fig. 8. A frame of speech is approximately 20 ms in length, and is composed of typically three to four subframes. The LPC excitation is transmitted only once per subframe, while the LPC coefficients are transmitted only once per frame. The LPC coefficients are computed by analyzing a window of speech that is usually longer than the speech frame (typically 30-60 ms). In order to minimize the number of future samples required to compute LPC coefficients, many LPC-AS coders use an asymmetric window that may include several hundred milliseconds of past context, but that emphasizes the samples of the current frame [21,84].

The perceptually weighted original signal $s_w(n)$ and weighted reconstructed signal $\hat{s}_w(n)$ in a given subframe



Figure 8. The frame/subframe structure of most LPC analysis by synthesis coders.

are often written as *L*-dimensional row vectors *S* and \hat{S} , where the dimension *L* is the length of a subframe:

$$S_w = [s_w(0), \dots, s_w(L-1)], \quad \hat{S}_w = [\hat{s}_w(0), \dots, \hat{s}_w(L-1)]$$
(21)

The core of an LPC-AS coder is the closed-loop search for an optimum coded excitation vector U, where U is typically composed of an "adaptive codebook" component representing the periodicity, and a "stochastic codebook" component representing the noiselike part of the excitation. In general, U may be represented as the weighted sum of several "shape vectors" X_m , m = $1, \ldots, M$, which may be drawn from several codebooks, including possibly multiple adaptive codebooks and multiple stochastic codebooks:

$$U = GX, \quad G = [g_1, g_2, \ldots], \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \end{bmatrix}$$
(22)

The choice of shape vectors and the values of the gains g_m are jointly optimized in a closed-loop search, in order to minimize the perceptually weighted error metric $|S_w - \hat{S}_w|^2$.

The value of S_w may be computed prior to any codebook search by perceptually weighting the input speech vector. The value of \hat{S}_w must be computed separately for each candidate excitation, by synthesizing the speech signal $\hat{s}(n)$, and then perceptually weighting to obtain $\hat{s}_w(n)$. These operations may be efficiently computed, as described below.

4.4.1. Zero State Response and Zero Input Response. Let the filter H(z) be defined as the composition of the LPC synthesis filter and the perceptual weighting filter, thus H(z) = W(z)/A(z). The computational complexity of the excitation parameter search may be greatly simplified if \hat{S}_w is decomposed into the zero input response (ZIR) and zero state response (ZSR) of H(z) [97]. Note that the weighted reconstructed speech signal is

$$\hat{S}_w = [\hat{s}_w(0), \dots, \hat{s}_w(L-1)], \quad \hat{s}_w(n) = \sum_{i=0}^{\infty} h(i)u(n-i)$$
(23)

where h(n) is the infinite-length impulse response of H(z). Suppose that $\hat{s}_w(n)$ has already been computed for n < 0, and the coder is now in the process of choosing the optimal u(n) for the subframe $0 \le n \le L - 1$. The sum above can be divided into two parts: a part that depends on the current subframe input, and a part that does not:

$$\hat{S}_w = \hat{S}_{\text{ZIR}} + UH \tag{24}$$

where \hat{S}_{ZIR} contains samples of the zero input response of H(z), and the vector UH contains the zero state response. The zero input response is usually computed by implementing the recursive filter H(z) = W(z)/A(z) as the sequence of two IIR filters, and allowing the two filters to run for L samples with zero input. The zero state response is usually computed as the matrix product UH, where

$$H = \begin{bmatrix} h(0) & h(1) & \dots & h(L-1) \\ 0 & h(0) & \dots & h(L-2) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & h(0) \end{bmatrix},$$
$$U = [u(0), \dots, u(L-1)]$$
(25)

Given a candidate excitation vector U, the perceptually weighted error vector E may be defined as

$$E_w = S_w - \hat{S}_w = \tilde{S} - UH \tag{26}$$

where the target vector \tilde{S} is

$$\tilde{S} = S_w - \hat{S}_{\text{ZIR}} \tag{27}$$

The target vector needs to be computed only once per subframe, prior to the codebook search. The objective of the codebook search, therefore, is to find an excitation vector U that minimizes $|\tilde{S} - UH|^2$.

4.4.2. Optimum Gain and Optimum Excitation. Recall that the excitation vector U is modeled as the weighted sum of a number of codevectors X_m , m = 1, ..., M. The perceptually weighted error is therefore:

$$|E|^{2} = |\tilde{S} - GXH|^{2} = \tilde{S}\tilde{S}' - 2GXH\tilde{S}' + GXH(GXH)' \quad (28)$$

where prime denotes transpose. Minimizing $|E|^2$ requires optimum choice of the shape vectors X and of the gains G. It turns out that the optimum gain for each excitation vector can be computed in closed form. Since the optimum gain can be computed in closed form, it need not be computed during the closed-loop search; instead, one can simply assume that each candidate excitation, if selected, would be scaled by its optimum gain. Assuming an optimum gain results in an extremely efficient criterion for choosing the optimum excitation vector [3].

Suppose we define the following additional bits of notation:

$$R_X = XHS', \quad \Sigma = XH(XH)'$$
 (29)

Then the mean-squared error is

$$|E|^2 = \tilde{S}\tilde{S}' - 2GR_X + G\Sigma G' \tag{30}$$

For any given set of shape vectors X, G is chosen so that $|E|^2$ is minimized, which yields

$$G = R'_X \Sigma^{-1} \tag{31}$$

If we substitute the minimum MSE value of G into Eq. (30), we get

$$|E|^2 = \tilde{S}\tilde{S}' - R'_X \Sigma^{-1} R_X \tag{32}$$

Hence, in order to minimize the perceptually weighted MSE, we choose the shape vectors X in order to maximize the covariance-weighted sum of correlations:

$$X_{\rm opt} = \arg\max(R'_X \Sigma^{-1} R_X) \tag{33}$$

When the shape matrix X contains more than one row, the matrix inversion in Eq. (33) is often computed using approximate algorithms [4]. In the VSELP coder [25], X is transformed using a modified Gram-Schmidt orthogonalization so that Σ has a diagonal structure, thus simplifying the computation of Eq. (33).

4.5. Types of LPC-AS Coder

4.5.1. Multipulse LPC (MPLPC). In the multipulse LPC algorithm [4,50], the shape vectors are impulses. U is typically formed as the weighted sum of 4-8 impulses per subframe.

The number of possible combinations of impulses grows exponentially in the number of impulses, so joint optimization of the positions of all impulses is usually impossible. Instead, most MPLPC coders optimize the pulse positions one at a time, using something like the following strategy. First, the weighted zero state response of H(z) corresponding to each impulse location is computed. If C_k is an impulse located at n = k, the corresponding weighted zero state response is

$$C_k H = [0, \dots, 0, h(0), h(1), \dots, h(L - k - 1)]$$
(34)

The location of the first impulse is chosen in order to optimally approximate the target vector $\tilde{S}_1 = \tilde{S}$, using the methods described in the previous section. After selecting the first impulse location k_1 , the target vector is updated according to

$$\tilde{S}_m = \tilde{S}_{m-1} - C_{k_{m-1}}H$$
 (35)

Additional impulses are chosen until the desired number of impulses is reached. The gains of all pulses may be reoptimized after the selection of each new pulse [87].

Variations are possible. The multipulse coder described in ITU standard G.723.1 transmits a single gain for all the impulses, plus sign bits for each individual impulse. The G.723.1 coder restricts all impulse locations to be either odd or even; the choice of odd or even locations is coded using one bit per subframe [50]. The regular pulse excited LPC algorithm, which was the first GSM full-rate speech coder, synthesized speech using a train of impulses spaced one per 4 samples, all scaled by a single gain term [65]. The alignment of the pulse train was restricted to one of four possible locations, chosen in a closed-loop fashion together with a gain, an adaptive codebook delay, and an adaptive codebook gain.

Singhal and Atal demonstrated that the quality of MPLPC may be improved at low bit rates by modeling the periodic component of an LPC excitation vector using a pitch prediction filter [87]. Using a pitch prediction filter, the LPC excitation signal becomes

$$u(n) = bu(n - D) + \sum_{m=1}^{M} c_{k_m}(n)$$
(36)

where the signal $c_k(n)$ is an impulse located at n = kand b is the pitch prediction filter gain. Singhal and Atal proposed choosing D before the locations of any impulses are known, by minimizing the following perceptually weighted error:

$$|E_D|^2 = |\tilde{S} - bX_D H|^2, X_D = [u(-D), \dots, u((L-1) - D)]$$
(37)

The G.723.1 multipulse LPC coder and the GSM (Global System for Mobile Communication) full-rate RPE-LTP (regular-pulse excitation with long-term prediction) coder both use a closed-loop pitch predictor, as do all standardized variations of the CELP coder (see Sections 4.5.2 and 4.5.3). Typically, the pitch delay and gain are optimized first, and then the gains of any additional excitation vectors (e.g., impulses in an MPLPC algorithm) are selected to minimize the remaining error.

4.5.2. Code-Excited LPC (CELP). LPC analysis finds a filter 1/A(z) whose excitation is uncorrelated for correlation distances smaller than the order of the filter. Pitch prediction, especially closed-loop pitch prediction, removes much of the remaining intersample correlation. The spectrum of the pitch prediction residual looks like the spectrum of uncorrelated Gaussian noise, but replacing the residual with real noise (noise that is independent of the original signal) yields poor speech quality. Apparently, some of the temporal details of the pitch prediction residual using a stochastic excitation vector $c_k(n)$ chosen from a list of stochastic excitation vectors, $k = 1, \ldots, K$, known to both the transmitter and receiver [85]:

$$u(n) = bu(n-D) + gc_k(n)$$
(38)

The list of stochastic excitation vectors is called a *stochastic codebook*, and the index of the stochastic codevector is chosen in order to minimize the perceptually weighted error metric $|E_k|^2$. Rose and Barnwell discussed the similarity between the search for an optimum predictor delay D [82], and the search for an optimum predictor delay D [82], and Kleijn et al. coined the term "adaptive codebook" to refer to the list of delayed excitation signals u(n-D) which the coder considers during closed-loop pitch delay optimization (Fig. 9).

The CELP algorithm was originally not considered efficient enough to be used in real-time speech coding, but a number of computational simplifications were proposed that resulted in real-time CELP-like algorithms. Trancoso and Atal proposed efficient search methods based on the truncated impulse response of the filter W(z)/A(z), as discussed in Section 4.4 [3,97]. Davidson and Lin separately proposed center clipping the stochastic codevectors, so that most of the samples in each codevector are zero [15,67]. Lin also proposed structuring the stochastic codebook so that each codevector is a slightly-shifted version of the previous codevector; such a codebook is called an overlapped codebook [67]. Overlapped stochastic codebooks are rarely used in practice today, but overlapped-codebook search methods are often used to reduce the computational complexity of an adaptive codebook search. In the search of



Figure 9. The code-excited LPC algorithm (CELP) constructs an LPC excitation signal by optimally choosing input vectors from two codebooks: an "adaptive" codebook, which represents the pitch periodicity; and a "stochastic" codebook, which represents the unpredictable innovations in each speech frame.

an overlapped codebook, the correlation R_X and autocorrelation Σ introduced in Section 4.4 may be recursively computed, thus greatly reducing the complexity of the codebook search [63].

Most CELP coders optimize the adaptive codebook index and gain first, and then choose a stochastic codevector and gain in order to minimize the remaining perceptually weighted error. If all the possible pitch periods are longer than one subframe, then the entire content of the adaptive codebook is known before the beginning of the codebook search, and the efficient overlapped codebook search methods proposed by Lin may be applied [67]. In practice, the pitch period of a female speaker is often shorter than one subframe. In order to guarantee that the entire adaptive codebook is known before beginning a codebook search, two methods are commonly used: (1) the adaptive codebook search may simply be constrained to only consider pitch periods longer than L samples — in this case, the adaptive codebook will lock onto values of D that are an integer multiple of the actual pitch period (if the same integer multiple is not chosen for each subframe, the reconstructed speech quality is usually good); and (2) adaptive codevectors with delays of D < L may be constructed by simply repeating the most recent D samples as necessary to fill the subframe.

4.5.3. SELP, **VSELP**, **ACELP**, **and LD-CELP**. Rose and Barnwell demonstrated that reasonable speech quality is achieved if the LPC excitation vector is computed completely recursively, using two closed-loop pitch predictors in series, with no additional information [82]. In their "self-excited LPC" algorithm (SELP), the LPC excitation is initialized during the first subframe using a vector of samples known at both the transmitter and receiver. For all frames after the first, the excitation is the sum of an arbitrary number of adaptive codevectors:

$$u(n) = \sum_{m=1}^{M} b_m u(n - D_m)$$
(39)

Kleijn et al. developed efficient recursive algorithms for searching the adaptive codebook in SELP coder and other LPC-AS coders [63].

Just as there may be more than one adaptive codebook, it is also possible to use more than one stochastic codebook. The vector-sum excited LPC algorithm (VSELP) models the LPC excitation vector as the sum of one adaptive and two stochastic codevectors [25]:

$$u(n) = bu(n - D) + \sum_{m=1}^{2} g_m c_{k_m}(n)$$
(40)

The two stochastic codebooks are each relatively small (typically 32 vectors), so that each of the codebooks may be searched efficiently. The adaptive codevector and the two stochastic codevectors are chosen sequentially. After selection of the adaptive codevector, the stochastic codebooks are transformed using a modified Gram–Schmidt orthogonalization, so that the perceptually weighted speech vectors generated during the first stochastic codebook search are all orthogonal to the perceptually weighted adaptive codevector. Because of this orthogonalization, the stochastic codebook search results in the choice of a stochastic codevector that is jointly optimal with the adaptive codevector, rather than merely sequentially optimal. VSELP is the basis of the Telecommunications Industry Associations digital cellular standard IS-54.

The algebraic CELP (ACELP) algorithm creates an LPC excitation by choosing just one vector from an adaptive codebook and one vector from a fixed codebook. In the ACELP algorithm, however, the fixed codebook is composed of binary-valued or trinary-valued algebraic codes, rather than the usual samples of a Gaussian noise process [1]. Because of the simplicity of the codevectors, it is possible to search a very large fixed codebook very quickly using methods that are a hybrid of standard CELP and MPLPC search algorithms. ACELP is the basis of the ITU standard G.729 coder at 8 kbps. ACELP codebooks may be somewhat larger than the codebooks in a standard CELP coder; the codebook in G.729, for example, contains 8096 codevectors per subframe.

Most LPC-AS coders operate at very low bit rates, but require relatively large buffering delays. The low-delay CELP coder (LD-CELP) operates at 16 kbps [10,47] and is designed to obtain the best possible speech quality, with the constraint that the total algorithmic delay of a tandem coder and decoder must be no more than 2 ms. LPC analysis and codevector search are computed once per 2 ms (16 samples). Transmission of LPC coefficients once per two milliseconds would require too many bits, so LPC coefficients are computed in a recursive backwardadaptive fashion. Before coding or decoding each frame, samples of $\hat{s}(n)$ from the previous frame are windowed, and used to update a recursive estimate of the autocorrelation function. The resulting autocorrelation coefficients are similar to those that would be obtained using a relatively long asymmetric analysis window. LPC coefficients are then computed from the autocorrelation function using the Levinson-Durbin algorithm.

4.6. Line Spectral Frequencies (LSFs) or Line Spectral Pairs (LSPs)

Linear prediction can be viewed as an inverse filtering procedure in which the speech signal is passed through an all-zero filter A(z). The filter coefficients of A(z) are chosen such that the energy in the output, that is, the residual or error signal, is minimized. Alternatively, the inverse filter A(z) can be transformed into two other filters P(z) and Q(z). These new filters turn out to have some interesting properties, and the representation based on them, called the *line spectrum pairs* [89,91], has been used in speech coding and synthesis applications.

Let A(z) be the frequency response of an LPC inverse filter of order p:

$$A(z) = -\sum_{i=0}^{p} a_i z^{-i}$$

with $a_0 = -1$. The a_i values are real, and all the zeros of A(z) are inside the unit circle.

If we use the lattice formulation of LPC, we arrive at a recursive relation between the *m*th stage $[A_m(z)]$ and the one before it $[A_{m-1}(z)]$ For the *p*th-order inverse filter, we have

$$A_p(z) = A_{p-1}(z) - k_p z^{-p} A_{p-1}(z^{-1})$$

By allowing the recursion to go one more iteration, we obtain

$$A_{p+1}(z) = A_p(z) - k_{p+1} z^{-(p+1)} A_p(z^{-1})$$
(41)

If we choose $k_{p+1} = \pm 1$ in Eq. (41), we can define two new polynomials as follows:

$$P(z) = A(z) - z^{-(p+1)}A(z^{-1})$$
(42)

$$Q(z) = A(z) + z^{-(p+1)}A(z^{-1})$$
(43)

Physically, P(z) and Q(z) can be interpreted as the inverse transfer function of the vocal tract for the *open-glottis* and *closed-glottis* boundary conditions, respectively [22], and P(z)/Q(z) is the driving-point impedance of the vocal tract as seen from the glottis [36].

If p is odd, the formulae for p_n and q_n are as follows:

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1})$$

=
$$\prod_{n=1}^{(p+1)/2} (1 - e^{jp_n}z^{-1})(1 - e^{-jp_n}z^{-1})$$
(44)

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1})$$

= $(1 - z^{-2}) \prod_{n=1}^{(p-1)/2} (1 - e^{jq_n} z^{-1})(1 - e^{-jq_n} z^{-1})$ (45)

The LSFs have some interesting characteristics: the frequencies $\{p_n\}$ and $\{q_n\}$ are related to the formant frequencies; the dynamic range of $\{p_n\}$ and $\{q_n\}$ is limited and the two alternate around the unit circle $(0 \le p_1 \le q_1 \le p_2 \ldots); \{p_n\}$ and $\{q_n\}$ are correlated so that intraframe prediction is possible; and they change slowly from one frame to another, hence, interframe prediction is also possible. The interleaving nature of the $\{p_n\}$ and $\{q_n\}$ allow for efficient iterative solutions [58].

Almost all LPC-based coders today use the LSFs to represent the LP parameters. Considerable recent research has been devoted to methods for efficiently quantizing the LSFs, especially using vector quantization (VQ) techniques. Typical algorithms include predictive VQ, split VQ [76], and multistage VQ [66,74]. All of these methods are used in the ITU standard ACELP coder G.729: the moving-average vector prediction residual is quantized using a 7-bit first-stage codebook, followed by second-stage quantization of two subvectors using independent 5-bit codebooks, for a total of 17 bits per frame [49,84].

5. LPC VOCODERS

5.1. The LPC-10e Vocoder

The 2.4-kbps LPC-10e vocoder (Fig. 10) is one of the earliest and one of the longest-lasting standards for lowbit-rate digital speech coding [8,16]. This standard was originally proposed in the 1970s, and was not officially replaced until the selection of the MELP 2.4-kbps coding standard in 1996 [64]. Speech coded using LPC-10e sounds metallic and synthetic, but it is intelligible.

In the LPC-10e algorithm, speech is first windowed using a Hamming window of length 22.5ms. The gain (G) and coefficients (a_i) of a linear prediction filter are calculated for the entire frame using the Levinson-Durbin recursion. Once G and a_i have been computed, the LPC residual signal d(n) is computed:

$$d(n) = \frac{1}{G} \left(s(n) - \sum_{i=1}^{p} a_i s(n-i) \right)$$
(46)

The residual signal d(n) is modeled using either a periodic train of impulses (if the speech frame is voiced) or an uncorrelated Gaussian random noise signal (if the frame is unvoiced). The voiced/unvoiced decision is based on the average magnitude difference function (AMDF),

$$\Phi_d(m) = \frac{1}{N - |m|} \sum_{n = |m|}^{N - 1} |d(n) - d(n - |m|)|$$
(47)

The frame is labeled as voiced if there is a trough in $\Phi_d(m)$ that is large enough to be caused by voiced excitation. Only values of *m* between 20 and 160 are examined, corresponding to pitch frequencies between 50 and 400 Hz. If the minimum value of $\Phi_d(m)$ in this range is less than



Figure 10. A simplified model of speech production whose parameters can be transmitted efficiently across a digital channel.

a threshold, the frame is declared voiced, and otherwise it is declared unvoiced [8].

If the frame is voiced, then the LPC residual is represented using an impulse train of period T_0 , where

$$T_0 = \arg\min_{m=20}^{160} \Phi_d(m)$$
(48)

If the frame is unvoiced, a pitch period of $T_0 = 0$ is transmitted, indicating that an uncorrelated Gaussian random noise signal should be used as the excitation of the LPC synthesis filter.

5.2. Mixed-Excitation Linear Prediction (MELP)

The mixed-excitation linear prediction (MELP) coder [69] was selected in 1996 by the United States Department of Defense Voice Processing Consortium (DDVPC) to be the U.S. Federal Standard at 2.4 kbps, replacing LPC-10e. The MELP coder is based on the LPC model with additional features that include mixed excitation, aperiodic pulses, adaptive spectral enhancement, pulse dispersion filtering, and Fourier magnitude modeling [70]. The synthesis model for the MELP coder is illustrated in Fig. 11. LP coefficients are converted to LSFs and a multistage vector quantizer (MSVQ) is used to quantize the LSF vectors. For voiced segments a total of 54 bits that represent: LSF parameters (25), Fourier magnitudes of the prediction residual signal (8), gain (8), pitch (7), bandpass voicing (4), aperiodic flag (1), and a sync bit are sent. The Fourier magnitudes are coded with an 8-bit VQ and the associated codebook is searched with a perceptuallyweighted Euclidean distance. For unvoiced segments, the Fourier magnitudes, bandpass voicing, and the aperiodic flag bit are not sent. Instead, 13 bits that implement forward error correction (FEC) are sent. The performance of MELP at 2.4 kbps is similar to or better than that of the federal standard at 4.8 kbps (FS 1016) [92]. Versions of MELP coders operating at 1.7 kbps [68] and 4.0 kbps [90] have been reported.

5.3. Multiband Excitation (MBE)

In multiband excitation (MBE) coding the voiced/unvoiced decision is not a binary one; instead, a series of voicing decisions are made for independent harmonic intervals [31]. Since voicing decisions can be made in different frequency bands individually, synthesized speech may be partially voiced and partially unvoiced. An improved version of the MBE was introduced in the late 1980s [7,35] and referred to as the IMBE coder. The IMBE



Figure 11. The MELP speech synthesis model.

at 2.4 kbps produces better sound quality than does the LPC-10e. The IMBE was adopted as the Inmarsat-M coding standard for satellite voice communication at a total rate of 6.4 kbps, including 4.15 kbps of source coding and 2.25 kbps of channel coding [104]. The Advanced MBE (AMBE) coder was adopted as the Inmarsat Mini-M standard at a 4.8 kbps total data rate, including 3.6 kbps of speech and 1.2 kbps of channel coding [18,27]. In [14] an enhanced multiband excitation (EMBE) coder was presented. The distinguishing features of the EMBE coder include signal-adaptive multimode spectral modeling and parameter quantization, a two-band signal-adaptive frequency-domain voicing decision, a novel VQ scheme for the efficient encoding of the variable-dimension spectral magnitude vectors at low rates, and multiclass selective protection of spectral parameters from channel errors. The 4-kbps EMBE coder accounts for both source (2.9 kbps) and channel (1.1 kbps) coding and was designed for satellitebased communication systems.

5.4. Prototype Waveform Interpolative (PWI) Coding

A different kind of coding technique that has properties of both waveform and LPC-based coders has been proposed [59,60] and is called prototype waveform interpolation (PWI). PWI uses both interpolation in the frequency domain and forward-backward prediction in the time domain. The technique is based on the assumption that, for voiced speech, a perceptually accurate speech signal can be reconstructed from a description of the waveform of a single, representative pitch cycle per interval of 20–30 ms. The assumption exploits the fact that voiced speech can be interpreted as a concentration of slowly evolving pitch cycle waveforms. The prototype waveform is described by a set of linear prediction (LP) filter coefficients describing the formant structure and a prototype excitation waveform, quantized with analysis-by-synthesis procedures. The speech signal is reconstructed by filtering an excitation signal consisting of the concatenation of (infinitesimal) sections of the instantaneous excitation waveforms. By coding the voiced and unvoiced components separately, a 2.4-kbps version of the coder performed similarly to the 4.8-kbps FS1016 standard [61].

Recent work has aimed at reducing the computational complexity of the coder for rates between 1.2 and 2.4 kbps by including a time-varying waveform sampling rate and a cubic B-spline waveform representation [62,86].

6. MEASURES OF SPEECH QUALITY

Deciding on an appropriate measurement of quality is one of the most difficult aspects of speech coder design, and is an area of current research and standardization. Early military speech coders were judged according to only one criterion: intelligibility. With the advent of consumergrade speech coders, intelligibility is no longer a sufficient condition for speech coder acceptability. Consumers want speech that sounds "natural." A large number of subjective and objective measures have been developed to quantify "naturalness," but it must be stressed that any scalar measurement of "naturalness" is an oversimplification. "Naturalness" is a multivariate quantity, including such factors as the metallic versus breathy quality of speech, the presence of noise, the color of the noise (narrowband noise tends to be more annoying than wideband noise, but the parameters that predict "annoyance" are not well understood), the presence of unnatural spectral envelope modulations (e.g., flutter noise), and the absence of natural spectral envelope modulations.

6.1. Psychophysical Measures of Speech Quality (Subjective Tests)

The final judgment of speech coder quality is the judgment made by human listeners. If consumers (and reviewers) like the way the product sounds, then the speech coder is a success. The reaction of consumers can often be predicted to a certain extent by evaluating the reactions of experimental listeners in a controlled psychophysical testing paradigm. Psychophysical tests (often called "subjective tests") vary depending on the quantity being evaluated, and the structure of the test.

6.1.1. Intelligibility. Speech coder intelligibility is evaluated by coding a number of prepared words, asking listeners to write down the words they hear, and calculating the percentage of correct transcriptions (an adjustment for guessing may be subtracted from the score). The diagnostic rhyme test (DRT) and diagnostic alliteration test (DALT) are intelligibility tests which use a controlled vocabulary to test for specific types of intelligibility loss [101,102]. Each test consists of 96 pairs of confusable words spoken in isolation. The words in a pair differ in only one distinctive feature, where the distinctive feature dimensions proposed by Voiers are voicing, nasality, sustention, sibilation, graveness, and compactness. In the DRT, the words in a pair differ in only one distinctive feature of the initial consonant; for instance, "jest" and "guest" differ in the sibilation of the initial consonant. In the DALT, words differ in the final consonant; for instance, "oaf" and "oath" differ in the graveness of the final consonant. Listeners hear one of the words in each pair, and are asked to select the word from two written alternatives. Professional testing firms employ trained listeners who are familiar with the speakers and speech tokens in the database, in order to minimize test-retest variability.

Intelligibility scores quoted in the speech coding literature often refer to the composite results of a DRT. In a comparison of two federal standard coders, the LPC 10e algorithm resulted in 90% intelligibility, while the FS-1016 CELP algorithm had 91% intelligibility [64]. An evaluation of waveform interpolative (WI) coding published DRT scores of 87.2% for the WI algorithm, and 87.7% for FS-1016 [61].

6.1.2. Numerical Measures of Perceptual Quality. Perhaps the most commonly used speech quality measure is the mean opinion score (MOS). A mean opinion score is computed by coding a set of spoken phrases using a variety of coders, presenting all of the coded speech together with undegraded speech in random order, asking listeners to rate the quality of each phrase on a numerical scale, and then averaging the numerical

ratings of all phrases coded by a particular coder. The five-point numerical scale is associated with a standard set of descriptive terms: 5 = excellent, 4 = good, 3 = fair, 2 = poor, and 1 = bad. A rating of 4 is supposed to correspond to standard toll-quality speech, quantized at 64 kbps using ITU standard G.711 [48].

Mean opinion scores vary considerably depending on background noise conditions; for example, CVSD performs significantly worse than LPC-based methods in quiet recording conditions, but significantly better under extreme noise conditions [96]. Gender of the speaker may also affect the relative ranking of coders [96]. Expert listeners tend to give higher rankings to speech coders with which they are familiar, even when they are not consciously aware of the order in which coders are presented [96]. Factors such as language and location of the testing laboratory may shift the scores of all coders up or down, but tend not to change the rank order of individual coders [39]. For all of these reasons, a serious MOS test must evaluate several reference coders in parallel with the coder of interest, and under identical test conditions. If an MOS test is performed carefully, intercoder differences of approximately 0.15 opinion points may be considered significant. Figure 12 is a plot of MOS as a function of bit rate for coders evaluated under quiet listening conditions in five published studies (one study included separately tabulated data from two different testing sites [96]).

The diagnostic acceptability measure (DAM) is an attempt to control some of the factors that lead to variability in published MOS scores [100]. The DAM employs trained listeners, who rate the quality of standardized test phrases on 10 independent perceptual scales, including six scales that rate the speech itself (fluttering, thin, rasping, muffled, interrupted, nasal), and four scales that rate the background noise (hissing, buzzing, babbling, rumbling). Each of these is a 100point scale, with a range of approximately 30 points between the LPC-10e algorithm (50 points) and clean speech (80 points) [96]. Scores on the various perceptual scales are combined into a composite quality rating. DAM scores are useful for pointing out specific defects in a speech coding algorithm. If the only desired test outcome is a relative quality ranking of multiple coders, a carefully controlled MOS test in which all coders of interest are tested under the same conditions may be as reliable as DAM testing [96].

6.1.3. Comparative Measures of Perceptual Quality. It is sometimes difficult to evaluate the statistical significance of a reported MOS difference between two coders. A more powerful statistical test can be applied if coders are evaluated in explicit A/B comparisons. In a comparative test, a listener hears the same phrase coded by two different coders, and chooses the one that sounds better. The result of a comparative test is an apparent preference score, and an estimate of the significance of the observed preference; for example, in a 1999 study, WI coding at 4.0 kbps was preferred to 4 kbps HVXC 63.7% of the time, to 5.3 kbps G.723.1 57.5% of the time (statistically significant differences), and to 6.3 kbps G.723.1 53.9% of the time (not statistically significant) [29]. It should be



noted that "statistical significance" in such a test refers only to the probability that the same listeners listening to the same waveforms will show the same preference in a future test.

6.2. Algorithmic Measures of Speech Quality (Objective Measures)

Psychophysical testing is often inconvenient; it is not possible to run psychophysical tests to evaluate every proposed adjustment to a speech coder. For this reason, a number of algorithms have been proposed that approximate, to a greater or lesser extent, the results of psychophysical testing.

The signal-to-noise ratio of a frame of N speech samples starting at sample number n may be defined as

$$SNR(n) = \frac{\sum_{m=n}^{n+N-1} s^2(m)}{\sum_{m=n}^{n+N-1} e^2(m)}$$
(49)

High-energy signal components can mask quantization error, which is synchronous with the signal component, or separated by at most a few tens of milliseconds. Over longer periods of time, listeners accumulate a general perception of quantization noise, which can be modeled as the average log segmental SNR:

SEGSNR =
$$\frac{1}{K} \sum_{k=0}^{K-1} 10 \log_{10} \text{SNR}(kN)$$
 (50)

Figure 12. Mean opinion scores from five published studies in quiet recording conditions — Jarvinen [53]. Kohler [64], MPEG [39], Yeldener [107], and the COMSAT and MPC sites from Tardelli et al. [96]: (A) unmodified speech, (B) ITU G.722 subband ADPCM, (C) ITU G.726 ADPCM, (D) ISO MPEG-II layer 3 subband audio coder, (E) DDVPC CVSD, (F) GSM full-rate RPE-LTP, (G) GSM EFR ACELP, (H) ITU G.729 ACELP, (I) TIA IS54 VSELP, (J) ITU G.723.1 MPLPC, (K) DDVPC FS-1016 CELP, (L) sinusoidal transform coding, (M) ISO MPEG-IV HVXC, (N) Inmarsat mini-M AMBE, (O) DDVPC FS-1015 LPC-10e, (P) DDVPC MELP.

High-amplitude signal components tend to mask quantization error components at nearby frequencies and times. A high-amplitude spectral peak in the speech signal is able to mask quantization error components at the same frequency, at higher frequencies, and to a much lesser extent, at lower frequencies. Given a short-time speech spectrum $S(e^{j\omega})$, it is possible to compute a short-time "masking spectrum" $M(e^{j\omega})$ which describes the threshold energy at frequency ω below which noise components are inaudible. The perceptual salience of a noise signal e(n)may be estimated by filtering the noise signal into Kdifferent subband signals $e_k(n)$, and computing the ratio between the noise energy and the masking threshold in each subband:

$$NMR(n,k) = \frac{\sum_{m=n}^{n+N-1} e_k^2(m)}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2 \, d\omega}$$
(51)

where ω_k is the lower edge of band k, and ω_{k+1} is the upper band edge. The band edges must be close enough together that all of the signal components in band k are effective in masking the signal $e_k(n)$. The requirement of effective masking is met if each band is exactly one Bark in width, where the Bark frequency scale is described in many references [71,77].

Fletcher has shown that the perceived loudness of a signal may be approximated by adding the cube roots of the signal power in each one-bark subband, after properly accounting for masking effects [20]. The total

loudness of a quantization noise signal may therefore be approximated as

$$NMR(n) = \sum_{k=0}^{K-1} \left(\frac{\sum_{m=n}^{n+N-1} e_k^2[m]}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2 \, d\omega} \right)^{1/3}$$
(52)

The ITU perceptual speech quality measure (PSQM) computes the perceptual quality of a speech signal by filtering the input and quantized signals using a Barkscale filterbank, nonlinearly compressing the amplitudes in each band, and then computing an average subband signal to noise ratio [51]. The development of algorithms that accurately predict the results of MOS or comparative testing is an area of active current research, and a number of improvements, alternatives, and/or extensions to the PSQM measure have been proposed. An algorithm that has been the focus of considerable research activity is the Bark spectral distortion measure [73,103,105,106]. The ITU has also proposed an extension of the PSQM standard called perceptual evaluation of speech quality (PESQ) [81], which will be released as ITU standard P.862.

7. NETWORK ISSUES

7.1. Voice over IP

Speech coding for the voice over Internet Protocol (VOIP) application is becoming important with the increasing dependency on the Internet. The first VoIP standard was published in 1998 as recommendation H.323 [52] by the International Telecommunications Union (ITU-T). It is a protocol for multimedia communications over local area networks using packet switching, and the voice-only subset of it provides a platform for IP-based telephony. At high bit rates, H.323 recommends the coders G.711 (3.4 kHz at 48, 56, and 64 kbps) and G.722 (wideband speech and music at 7 kHz operating at 48, 56, and 64 kbps) while at the lower bit rates G.728 (3.4 kHz at 16 kbps), G.723 (5.3 and 6.5 kbps), and G.729 (8 kbps) are recommended [52].

In 1999, a competing and simpler protocol named the Session Initiation Protocol (SIP) was developed by the Internet Engineering Task Force (IETF) Multiparty Multimedia Session Control working group and published as RFC 2543 [19]. SIP is a signaling protocol for Internet conferencing and telephony, is independent of the packet layer, and runs over UDP or TCP although it supports more protocols and handles the associations between Internet end systems. For now, both systems will coexist but it is predicted that the H.323 and SIP architectures will evolve such that two systems will become more similar.

Speech transmission over the Internet relies on sending "packets" of the speech signal. Because of network congestion, packet loss can occur, resulting in audible artifacts. High-quality VOIP, hence, would benefit from variable-rate source and channel coding, packet loss concealment, and jitter buffer/delay management. These are challenging issues and research efforts continue to generate high-quality speech for VOIP applications [38].

7.2. Embedded and Multimode Coding

When channel quality varies, it is often desirable to adjust the bit rate of a speech coder in order to match the channel capacity. Varying bit rates are achieved in one of two ways. In multimode speech coding, the transmitter and the receiver must agree on a bit rate prior to transmission of the coded bits. In embedded source coding, on the other hand, the bitstream of the coder operating at low bit rates is embedded in the bitstream of the coder operating at higher rates. Each increment in bit rate provides marginal improvement in speech quality. Lower bit rate coding is obtained by puncturing bits from the higher rate coder and typically exhibits graceful degradation in quality with decreasing bit rates.

ITU Standard G.727 describes an embedded ADPCM coder, which may be run at rates of 40, 32, 24, or 16 kbps (5, 4, 3, or 2 bits per sample) [46]. Embedded ADPCM algorithms are a family of variable bit rate coding algorithms operating on a sample per sample basis (as opposed to, e.g., a subband coder that operates on a frameby-frame basis) that allows for bit dropping after encoding. The decision levels of the lower-rate quantizers are subsets of those of the quantizers at higher rates. This allows for bit reduction at any point in the network without the need of coordination between the transmitter and the receiver.

The prediction in the encoder is computed using a more coarse quantization of $\hat{d}(n)$ than the quantization actually transmitted. For example, 5 bits per sample may be transmitted, but as few as 2 bits may be used to reconstruct $\hat{d}(n)$ in the prediction loop. Any bits not used in the prediction loop are marked as "optional" by the signaling channel mode flag. If network congestion disrupts traffic at a router between sender and receiver, the router is allowed to drop optional bits from the coded speech packets.

Embedded ADPCM algorithms produce codewords that contain enhancement and core bits. The feedforward (FF) path of the codec utilizes both enhancement bits and core bits, while the feedback (FB) path uses core bits only. With this structure, enhancement bits can be discarded or dropped during network congestion.

An important example of a multimode coder is QCELP, the speech coder standard that was adopted by the TIA North American digital cellular standard based on codedivision multiple access (CDMA) technology [9]. The coder selects one of four data rates every 20 ms depending on the speech activity; for example, background noise is coded at a lower rate than speech. The four rates are approximately 1 kbps (eighth rate), 2 kbps (quarter rate), 4 kbps (half rate), and 8 kbps (full rate). QCELP is based on the CELP structure but integrates implementation of the different rates, thus reducing the average bit rate. For example, at the higher rates, the LSP parameters are more finely quantized and the pitch and codebook parameters are updated more frequently [23]. The coder provides good quality speech at average rates of 4 kbps.

Another example of a multimode coder is ITU standard G.723.1, which is an LPC-AS coder that can operate at

2 rates: 5.3 or 6.3 kbps [50]. At 6.3 kbps, the coder is a multipulse LPC (MPLPC) coder while the 5.3-kbps coder is an algebraic CELP (ACELP) coder. The frame size is 30 ms with an additional lookahead of 7.5 ms, resulting in a total algorithmic delay of 67.5 ms. The ACELP and MPLPC coders share the same LPC analysis algorithm and frame/subframe structure, so that most of the program code is used by both coders. As mentioned earlier, in ACELP, an algebraic transformation of the transmitted index produces the excitation signal for the synthesizer. In MPLPC, on the other hand, minimizing the perceptualerror weighting is achieved by choosing the amplitude and position of a number of pulses in the excitation signal. Voice activity detection (VAD) is used to reduce the bit rate during silent periods, and switching from one bit rate to another is done on a frame-by-frame basis.

Multimode coders have been proposed over a wide variety of bandwidths. Taniguchi et al. proposed a multimode ADPCM coder at bit rates between 10 and 35 kbps [94]. Johnson and Taniguchi proposed a multimode CELP algorithm at data rates of 4.0-5.3 kbps in which additional stochastic codevectors are added to the LPC excitation vector when channel conditions are sufficiently good to allow high-quality transmission [55]. The European Telecommunications Standards Institute (ETSI) has recently proposed a standard for adaptive multirate coding at rates between 4.75 and 12.2 kbps.

7.3. Joint Source-Channel Coding

In speech communication systems, a major challenge is to design a system that provides the best possible speech quality throughout a wide range of channel conditions. One solution consists of allowing the transceivers to monitor the state of the communication channel and to dynamically allocate the bitstream between source and channel coding accordingly. For low-SNR channels, the source coder operates at low bit rates, thus allowing powerful forward error control. For high-SNR channels, the source coder uses its highest rate, resulting in high speech quality, but with little error control. An adaptive algorithm selects a source coder and channel coder based on estimates of channel quality in order to maintain a constant total data rate [95]. This technique is called adaptive multirate (AMR) coding, and requires the simultaneous implementation of an AMR source coder [24], an AMR channel coder [26,28], and a channel quality estimation algorithm capable of acquiring information about channel conditions with a relatively small tracking delay.

The notion of determining the relative importance of bits for further unequal error protection (UEP) was pioneered by Rydbeck and Sundberg [83]. Ratecompatible channel codes, such as Hagenauer's rate compatible punctured convolutional codes (RCPC) [34], are a collection of codes providing a family of channel coding rates. By puncturing bits in the bitstream, the channel coding rate of RCPC codes can be varied instantaneously, providing UEP by imparting on different segments different degrees of protection. Cox et al. [13] address the issue of channel coding and illustrate how RCPC codes can be used to build a speech transmission scheme for mobile radio channels. Their approach is based on a subband coder with dynamic bit allocation proportional to the average energy of the bands. RCPC codes are then used to provide UEP.

Relatively few AMR systems describing source and channel coding have been presented. The AMR systems [99,98,75,44] combine different types of variable rate CELP coders for source coding with RCPC and cyclic redundancy check (CRC) codes for channel coding and were presented as candidates for the European Telecommunications Standards Institute (ETSI) GSM AMR codec standard. In [88], UEP is applied to perceptually based audio coders (PAC). The bitstream of the PAC is divided into two classes and punctured convolutional codes are used to provide different levels of protection, assuming a BPSK constellation.

In [5,6], a novel UEP channel encoding scheme is introduced by analyzing how symbol-wise puncturing of symbols in a trellis code and the rate-compatibility constraint (progressive puncturing pattern) can be used to derive rate-compatible punctured trellis codes (RCPT). While conceptually similar to RCPC codes, RCPT codes are specifically designed to operate efficiently on large constellations (for which Euclidean and Hamming distances are no longer equivalent) by maximizing the residual Euclidean distance after symbol puncturing. Large constellation sizes, in turn, lead to higher throughput and spectral efficiency on high SNR channels. An AMR system is then designed based on a perceptually-based embedded subband encoder. Since perceptually based dynamic bit allocations lead to a wide range of bit error sensitivities (the perceptually least important bits being almost insensitive to channel transmission errors), the channel protection requirements are determined accordingly. The AMR systems utilize the new rate-compatible channel coding technique (RCPT) for UEP and operate on an 8-PSK constellation. The AMR-UEP system is bandwidth efficient, operates over a wide range of channel conditions and degrades gracefully with decreasing channel quality.

Systems using AMR source and channel coding are likely to be integrated in future communication systems since they have the capability for providing graceful speech degradation over a wide range of channel conditions.

8. STANDARDS

Standards for landline public switched telephone service (PSTN) networks are established by the International Telecommunication Union (ITU) (http://www.itu.int). The ITU has promulgated a number of important speech and waveform coding standards at high bit rates and with very low delay, including G.711 (PCM), G.727 and G.726 (ADPCM), and G.728 (LDCELP). The ITU is also involved in the development of internetworking standards, including the voice over IP standard H.323. The ITU has developed one widely used low-bit-rate coding standard (G.729), and a number of embedded and multimode speech coding standards operating at rates between 5.3 kbps (G.723.1) and 40 kbps (G.727). Standard G.729 is a speech coder operating at 8 kbps, based on algebraic code-excited LPC (ACELP) [49,84]. G.723.1 is a multimode coder, capable of operating at either 5.3 or 6.3 kbps [50]. G.722 is a standard for wideband speech coding, and the ITU will announce an additional wideband standard within the near future. The ITU has also published standards for the objective estimation of perceptual speech quality (P.861 and P.862).

The ITU is a branch of the International Standards Organization (ISO) (*http://www.iso.ch*). In addition to ITU activities, the ISO develops standards for the Moving Picture Experts Group (MPEG). The MPEG-2 standard included digital audiocoding at three levels of complexity, including the layer 3 codec commonly known as MP3 [72]. The MPEG-4 motion picture standard includes a structured audio standard [40], in which speech and audio "objects" are encoded with header information specifying the coding algorithm. Low-bit-rate speech coding is performed using harmonic vector excited coding (HVXC) [43] or code-excited LPC (CELP) [41], and audiocoding is performed using time-frequency coding [42]. The MPEG homepage is at *drogo.cselt.stet.it/mpeg*.

Standards for cellular telephony in Europe are established by the European Telecommunications Standards Institute (ETSI) (*http://www.etsi.org*). ETSI speech coding standards are published by the Global System for Mobile Telecommunications (GSM) subcommittee. All speech coding standards for digital cellular telephone use are based on LPC-AS algorithms. The first GSM standard coder was based on a precursor of CELP called *regular-pulse excitation with long-term prediction* (RPE-LTP) [37,65]. Current GSM standards include the enhanced full-rate codec GSM 06.60 [32,53] and the adaptive multirate codec [33]; both standards use algebraic code-excited LPC (ACELP). At the time of writing, both ITU and ETSI are expected to announce new standards for wideband speech coding in the near future. ETSI's standard will be based on GSM AMR.

The Telecommunications Industry Association (*http://www.tiaonline.org*) published some of the first U.S. digital cellular standards, including the vector-sum-excited LPC (VSELP) standard IS54 [25]. In fact, both the initial U.S. and Japanese digital cellular standards were based on the VSELP algorithm. The TIA has been active in the development of standard TR41 for voice over IP.

The U.S. Department of Defense Voice Processing Consortium (DDVPC) publishes speech coding standards for U.S. government applications. As mentioned earlier, the original FS-1015 LPC-10e standard at 2.4 kbps [8,16], originally developed in the 1970s, was replaced in 1996 by the newer MELP standard at 2.4 kbps [92]. Transmission at slightly higher bit rates uses the FS-1016 CELP (CELP) standard at 4.8 kbps [17,56,57]. Waveform applications use the continuously variable slope delta modulator (CVSD) at 16 kbps. Descriptions of all DDVPC standards and code for most are available at http://www.plh.af.mil/ddvpc/index.html.

9. FINAL REMARKS

In this article, we presented an overview of coders that compress speech by attempting to match the time waveform as closely as possible (waveform coders), and coders that attempt to preserve perceptually relevant spectral properties of the speech signal (LPC-based and subband coders). LPC-based coders use a speech production model to parameterize the speech signal, while subband coders filter the signal into frequency bands and assign bits by either an energy or perceptual criterion. Issues pertaining to networking, such as voice over IP and joint source-channel coding, were also touched on. There are several other coding techniques that we have not discussed in this article because of space limitations. We hope to have provided the reader with an overview of the fundamental techniques of speech compression.

Acknowledgments

This research was supported in part by the NSF and HRL. We thank Alexis Bernard and Tomohiko Taniguchi for their suggestions on earlier drafts of the article.

BIOGRAPHIES

Mark A. Hasegawa-Johnson received his S.B., S.M., and Ph.D. degrees in electrical engineering and computer science from MIT in 1989, 1989, and 1996, respectively. From 1989 to 1990 he worked as a research engineer at Fujitsu Laboratories Ltd., Kawasaki, Japan, where he developed and patented a multimodal CELP speech coder with an efficient algebraic fixed codebook. From 1996–1999 he was a postdoctoral fellow in the Electrical Engineering Department at UCLA. Since 1999, he has been on the faculty of the University of Illinois at Urbana-Champaign. Dr. Hasegawa-Johnson holds four U.S. patents and is the author of four journal articles and twenty conference papers. His areas of interest include speech coding, automatic speech understanding, acoustics, and the physiology of speech production.

Abeer Alwan received her Ph.D. in electrical engineering from MIT in 1992. Since then, she has been with the Electrical Engineering Department at UCLA, California, as an assistant professor (1992-1996), associate professor (1996-2000), and professor (2000-present). Professor Alwan established and directs the Speech Processing and Auditory Perception Laboratory at UCLA (http://www.icsl.ucla.edu/~spapl). Her research interests include modeling human speech production and perception mechanisms and applying these models to speechprocessing applications such as automatic recognition, compression, and synthesis. She is the recipient of the NSF Research Initiation Award (1993), the NIH FIRST Career Development Award (1994), the UCLA-TRW Excellence in Teaching Award (1994), the NSF Career Development Award (1995), and the Okawa Foundation Award in Telecommunications (1997). Dr. Alwan is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served as an elected member on the Acoustical Society of America Technical Committee on Speech Communication (1993-1999), on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics (1996-2000) and Speech Processing (1996-2001). She is an editor in chief of the journal Speech Communication.

SPEECH CODING: FUNDAMENTALS AND APPLICATIONS 2357

BIBLIOGRAPHY

- J.-P. Adoul, P. Mabilleau, M. Delprat, and S. Morisette, Fast CELP coding based on algebraic codes, *Proc. ICASSP*, 1987, pp. 1957–1960.
- B. S. Atal, Predictive coding of speech at low bit rates, *IEEE Trans. Commun.* 30: 600–614 (1982).
- B. S. Atal, High-quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders, *Proc. ICASSP*, 1986, pp. 1681–1684.
- 4. B. S. Atal and J. R. Remde, A new model of LPC excitation for producing natural-sounding speech at low bit rates, *Proc. ICASSP*, 1982, pp. 614–617.
- A. Bernard, X. Liu, R. Wesel, and A. Alwan, Channel adaptive joint-source channel coding of speech, Proc. 32nd Asilomar Conf. Signals, Systems, and Computers, 1998, Vol. 1, pp. 357–361.
- A. Bernard, X. Liu, R. Wesel, and A. Alwan, Embedded joint-source channel coding of speech using symbol puncturing of trellis codes, *Proc. IEEE ICASSP*, 1999, Vol. 5, pp. 2427–2430.
- M. S. Brandstein, P. A. Monta, J. C. Hardwick, and J. S. Lim, A real-time implementation of the improved MBE speech coder, *Proc. ICASSP*, 1990, Vol. 1: pp. 5–8.
- 8. J. P. Campbell and T. E. Tremain, Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm, *Proc. ICASSP*, 1986, pp. 473–476.
- 9. CDMA, Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard, Technical Report Proposed EIA/TIA Interim Standard, Telecommunications Industry Association TR45.5 Subcommittee, 1992.
- J.-H. Chen et al., A low delay CELP coder for the CCITT 16 kb/s speech coding standard, *IEEE J. Select. Areas Commun.* 10: 830-849 (1992).
- J.-H. Chen and A. Gersho, Adaptive postfiltering for quality enhancement of coded speech, *IEEE Trans. Speech Audio Process.* 3(1): 59-71 (1995).
- R. Cox et al., New directions in subband coding, *IEEE JSAC* 6(2): 391–409 (Feb. 1988).
- R. Cox, J. Hagenauer, N. Seshadri, and C. Sundberg, Subband speech coding and matched convolutional coding for mobile radio channels, *IEEE Trans. Signal Process.* **39**(8): 1717–1731 (Aug. 1991).
- A. Das and A. Gersho, Low-rate multimode multiband spectral coding of speech, *Int. J. Speech Tech.* 2(4): 317–327 (1999).
- G. Davidson and A. Gersho, Complexity reduction methods for vector excitation coding, *Proc. ICASSP*, 1986, pp. 2055–2058.
- DDVPC, LPC-10e Speech Coding Standard, Technical Report FS-1015, U.S. Dept. of Defense Voice Processing Consortium, Nov. 1984.
- DDVPC, CELP Speech Coding Standard, Technical Report FS-1016, U.S. Dept. of Defense Voice Processing Consortium, 1989.
- S. Dimolitsas, Evaluation of voice coded performance for the Inmarsat Mini-M system, Proc. 10th Int. Conf. Digital Satellite Communications, 1995.

- M. Handley et al., SIP: Session Initiation Protocol, IETF RFC, March 1999, http://www.cs.columbia.edu/hgs/sip/ sip.html.
- H. Fletcher, Speech and Hearing in Communication, Van Nostrand, Princeton, NJ, 1953.
- D. Florencio, Investigating the use of asymmetric windows in CELP vocoders, *Proc. ICASSP*, 1993, Vol. II, pp. 427–430.
- 22. S. Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, New York, 1989.
- 23. W. Gardner, P. Jacobs, and C. Lee, QCELP: A variable rate speech coder for CDMA digital cellular, in B. Atal, V. Cuperman, and A. Gersho, eds., *Speech and Audio Coding* for Wireless and Network Applications, Kluwer, Dordrecht, The Netherlands, 1993, pp. 85–93.
- 24. A. Gersho and E. Paksoy, An overview of variable rate speech coding for cellular networks, *IEEE Int. Conf. Selected Topics in Wireless Communications Proc.*, June 1999, pp. 172–175.
- I. Gerson and M. Jasiuk, Vector sum excited linear prediction (VSELP), in B. S. Atal, V. S. Cuperman, and A. Gersho, eds., Advances in Speech Coding, Kluwer, Dordrecht, The Netherlands, 1991, pp. 69–80.
- D. Goeckel, Adaptive coding for time-varying channels using outdated fading estimates, *IEEE Trans. Commun.* 47(6): 844–855 (1999).
- 27. R. Goldberg and L. Riek, A Practical Handbook of Speech Coders, CRC Press, Boca Raton, FL, 2000.
- A. Goldsmith and S. G. Chua, Variable-rate variable power MQAM for fading channels, *IEEE Trans. Commun.* 45(10): 1218-1230 (1997).
- O. Gottesman and A. Gersho, Enhanced waveform interpolative coding at 4 kbps, *IEEE Workshop on Speech Coding*, Piscataway, NY, 1999, pp. 90-92.
- K. Gould, R. Cox, N. Jayant, and M. Melchner, Robust speech coding for the indoor wireless channel, ATT Tech. J. 72(4): 64-73 (1993).
- D. W. Griffn and J. S. Lim, Multi-band excitation vocoder, *IEEE Trans. Acoust. Speech Signal Process.* 36(8): 1223-1235 (1988).
- 32. Special Mobile Group (GSM), Digital Cellular Telecommunications System: Enhanced Full Rate (EFR) Speech Transcoding, Technical Report GSM 06.60, European Telecommunications Standards Institute (ETSI), 1997.
- 33. Special Mobile Group (GSM), Digital Cellular Telecommunications System (Phase 2+): Adaptive Multi-rate (AMR) Speech Transcoding, Technical Report GSM 06.90, European Telecommunications Standards Institute (ETSI), 1998.
- J. Hagenauer, Rate-compatible punctured convolutional codes and their applications, *IEEE Trans. Commun.* 36(4): 389-400 (1988).
- J. C. Hardwick and J. S. Lim, A 4.8 kbps multi-band excitation speech coder, *Proc. ICASSP*, 1988, Vol. 1, pp. 374–377.
- M. Hasegawa-Johnson, Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model, J. Acoust. Soc. Am. 108(1): 457-460 (2000).
- 37. K. Hellwig et al., Speech codec for the european mobile radio system, *Proc. IEEE Global Telecomm. Conf.*, 1989.
- O. Hersent, D. Gurle, and J.-P. Petit, *IP Telephony*, Addison-Wesley, Reading, MA, 2000.

2358 SPEECH CODING: FUNDAMENTALS AND APPLICATIONS

- ISO, Report on the MPEG-4 Speech Codec Verification Tests, Technical Report JTC1/SC29/WG11, ISO/IEC, Oct. 1998.
- ISO/IEC, Information Technology—Coding of Audiovisual Objects, Part 3: Audio, Subpart 1: Overview, Technical Report ISO/JTC 1/SC 29/N2203, ISO/IEC, 1998.
- ISO/IEC, Information Technology—Coding of Audiovisual Objects, Part 3: Audio, Subpart 3: CELP, Technical Report ISO/JTC 1/SC 29/N2203CELP, ISO/IEC, 1998.
- ISO/IEC, Information Technology Coding of Audiovisual Objects, Part 3: Audio, Subpart 4: Time/Frequency Coding, Technical Report ISO/JTC 1/SC 29/N2203TF, ISO/IEC, 1998.
- ISO/IEC, Information Technology Very Low Bitrate Audio-Visual Coding, Part 3: Audio, Subpart 2: Parametric Coding, Technical Report ISO/JTC 1/SC 29/N2203PAR, ISO/IEC, 1998.
- 44. H. Ito, M. Serizawa, K. Ozawa, and T. Nomura, An adaptive multi-rate speech codec based on mp-celp coding algorithm for etsi amr standard, *Proc. ICASSP*, 1998, Vol. 1, pp. 137–140.
- ITU-T, 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM), Technical Report G.726, International Telecommunications Union, Geneva, 1990.
- 46. ITU-T, 5-, 4-, 3- and 2-bits per Sample Embedded Adaptive Differential Pulse Code Modulation (ADPCM), Technical Report G.727, International Telecommunications Union, Geneva, 1990.
- 47. ITU-T, Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction, Technical Report G.728, International Telecommunications Union, Geneva, 1992.
- ITU-T, Pulse Code Modulation (PCM) of Voice Frequencies, Technical Report G.711, International Telecommunications Union, Geneva, 1993.
- ITU-T, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), Technical Report G.729, International Telecommunications Union, Geneva, 1996.
- ITU-T, Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s, Technical Report G.723.1, International Telecommunications Union, Geneva, 1996.
- ITU-T, Objective Quality Measurement of Telephone-Band (300-3400 Hz) speech codecs, Technical Report P.861, International Telecommunications Union, Geneva, 1998.
- 52. ITU-T, Packet Based Multimedia Communications Systems, Technical Report H.323, International Telecommunications Union, Geneva, 1998.
- 53. K. Jarvinen et al., GSM enhanced full rate speech codec, *Proc. ICASSP*, 1997, pp. 771-774.
- 54. N. Jayant, J. Johnston, and R. Safranek, Signal compression based on models of human perception, *Proc. IEEE* 81(10): 1385-1421 (1993).
- 55. M. Johnson and T. Taniguchi, Low-complexity multi-mode VXC using multi-stage optimization and mode selection, *Proc. ICASSP*, 1991, pp. 221–224.
- 56. J. P. Campbell Jr., T. E. Tremain, and V. C. Welch, The DOD 4.8 KBPS standard (proposed federal standard 1016), in B. S. Atal, V. C. Cuperman, and A. Gersho, ed., Advances in Speech Coding, Kluwer, Dordrecht, The Netherlands, 1991, pp. 121–133.

- 57. J. P. Campbell, Jr., V. C. Welch, and T. E. Tremain, An expandable error-protected 4800 BPS CELP coder (U.S. federal standard 4800 BPS voice coder), *Proc. ICASSP*, 1989, 735–738.
- P. Kabal and R. Ramachandran, The computation of line spectral frequencies using chebyshev polynomials, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-34: 1419–1426 (1986).
- W. Kleijn, Speech coding below 4 kb/s using waveform interpolation, *Proc. GLOBECOM* 1991, Vol. 3, pp. 1879–1883.
- W. Kleijn and W. Granzow, Methods for waveform interpolation in speech coding, *Digital Signal Process.* 1(4): 215–230 (1991).
- W. Kleijn and J. Haagen, A speech coder based on decomposition of characteristic waveforms, *Proc. ICASSP*, 1995, pp. 508-511.
- W. Kleijn, Y. Shoham, D. Sen, and R. Hagen, A lowcomplexity waveform interpolation coder, *Proc. ICASSP*, 1996, pp. 212–215.
- W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, Improved speech quality and efficient vector quantization in SELP, *Proc. ICASSP*, 1988, pp. 155–158.
- M. Kohler, A comparison of the new 2400bps MELP federal standard with other standard coders, *Proc. ICASSP*, 1997, pp. 1587–1590.
- P. Kroon, E. F. Deprettere, and R. J. Sluyter, Regular-pulse excitation: A novel approach to effective and efficient multipulse coding of speech, *IEEE Trans. ASSP* 34: 1054–1063 (1986).
- W. LeBlanc, B. Bhattacharya, S. Mahmoud, and V. Cuperman, Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4kb/s speech coding, *IEEE Trans. Speech Audio Process.* 1: 373-385 (1993).
- 67. D. Lin, New approaches to stochastic coding of speech sources at very low bit rates, in I. T. Young et al., ed., *Signal Processing III: Theories and Applications*, Elsevier, Amsterdam, 1986, pp. 445-447.
- A. McCree and J. C. De Martin, A 1.7 kb/s MELP coder with improved analysis and quantization, *Proc. ICASSP*, 1998, Vol. 2, pp. 593-596.
- A. McCree et al., A 2.4 kbps MELP coder candidate for the new U.S. Federal standard, *Proc. ICASSP*, 1996, Vol. 1, pp. 200-203.
- 70. A. V. McCree and T. P. Barnwell, III, A mixed excitation LPC vocoder model for low bit rate speech coding, *IEEE Trans. Speech Audio Process.* 3(4): 242-250 (1995).
- B. C. J. Moore, An Introduction to the Psychology of Hearing, Academic Press, San Diego, (1997).
- P. Noll, MPEG digital audio coding, *IEEE Signal Process.* Mag. 14(5): 59–81 (1997).
- B. Novorita, Incorporation of temporal masking effects into bark spectral distortion measure, *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 665–668.
- E. Paksoy, W-Y. Chan, and A. Gersho, Vector quantization of speech LSF parameters with generalized product codes, *Proc. ICASSP*, 1992, pp. 33–36.
- E. Paksoy et al., An adaptive multi-rate speech coder for digital cellular telephony, *Proc. of ICASSP*, 1999, Vol. 1, pp. 193-196.

- K. K. Paliwal and B. S. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, *IEEE Trans. Speech Audio Process.* 1: 3-14 (1993).
- L. Rabiner and B.-H Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffis, NJ, 1993.
- L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.
- 79. R. P. Ramachandran and P. Kabal, Stability and performance analysis of pitch filters in speech coders, *IEEE Trans.* ASSP 35(7): 937–946 (1987).
- V. Ramamoorthy and N. S. Jayant, Enhancement of ADPCM speech by adaptive post-filtering, *AT&T Bell Labs. Tech. J.* 63(8): 1465-1475 (1984).
- A. Rix, J. Beerends, M. Hollier, and A. Hekstra, PESQ—the new ITU standard for end-to-end speech quality assessment, *AES 109th Convention*, Los Angeles, CA, Sept. 2000.
- R. C. Rose and T. P. Barnwell, III, The self-excited vocoder — an alternate approach to toll quality at 4800 bps, *Proc. ICASSP*, 1986, pp. 453-456.
- N. Rydbeck and C. E. Sundberg, Analysis of digital errors in non-linear PCM systems, *IEEE Trans. Commun.* COM-24: 59-65 (1976).
- 84. R. Salami et al., Design and description of CS-ACELP: A toll quality 8 kb/s speech coder, *IEEE Trans. Speech Audio Process.* 6(2): 116–130 (1998).
- M. R. Schroeder and B. S. Atal, Code-excited linear prediction (CELP): High-quality speech at very low bit rates, *Proc. ICASSP*, 1985, pp. 937–940.
- Y. Shoham, Very low complexity interpolative speech coding at 1.2 to 2.4 kbp, *Proc. ICASSP*, 1997, pp. 1599–1602.
- S. Singhal and B. S. Atal, Improving performance of multipulse LPC coders at low bit rates, *Proc. ICASSP*, 1984, pp. 1.3.1-1.3.4.
- D. Sinha and C.-E. Sundberg, Unequal error protection methods for perceptual audio coders, *Proc. ICASSP*, 1999, Vol. 5, pp. 2423–2426.
- F. Soong and B.-H. Juang, Line spectral pair (LSP) and speech data compression, *Proc. ICASSP*, 1984, pp. 1.10.1-1.10.4.
- J. Stachurski, A. McCree, and V. Viswanathan, High quality MELP coding at bit rates around 4 kb/s, *Proc. ICASSP*, 1999, Vol. 1, pp. 485–488.
- N. Sugamura and F. Itakura, Speech data compression by LSP speech analysis-synthesis technique, *Trans. IECE* J64-A(8): 599–606 (1981) (in Japanese).
- L. Supplee, R. Cohn, and J. Collura, MELP: The new federal standard at 2400 bps, *Proc. ICASSP*, 1997, pp. 1591–1594.
- B. Tang, A. Shen, A. Alwan, and G. Pottie, A perceptuallybased embedded subband speech coder, *IEEE Trans. Speech Audio Process.* 5(2): 131–140 (March 1997).
- T. Taniguchi, ADPCM with a multiquantizer for speech coding, *IEEE J. Select. Areas Commun.* 6(2): 410-424 (1988).
- T. Taniguchi, F. Amano, and S. Unagami, Combined source and channel coding based on multimode coding, *Proc. ICASSP*, 1990, pp. 477–480.
- 96. J. Tardelli and E. Kreamer, Vocoder intelligibility and quality test methods, *Proc. ICASSP*, 1996, pp. 1145-1148.

- 97. I. M. Trancoso and B. S. Atal, Efficient procedures for finding the optimum innovation in stochastic coders, *Proc. ICASSP*, 1986, pp. 2379–2382.
- A. Uvliden, S. Bruhn, and R. Hagen, Adaptive multi-rate. A speech service adapted to cellular radio network quality, *Proc. 32nd Asilomar Conf.*, 1998, Vol. 1, pp. 343-347.
- J. Vainio, H. Mikkola, K. Jarvinen, and P. Haavisto, GSM EFR based multi-rate codec family, *Proc. ICASSP*, 1998, Vol. 1, pp. 141–144.
- W. D. Voiers, Diagnostic acceptability measure for speech communication systems, Proc. ICASSP, 1977, pp. 204–207.
- W. D. Voiers, Evaluating processed speech using the diagnostic rhyme test, Speech Technol. 1(4): 30–39 (1983).
- 102. W. D. Voiers, Effects of noise on the discriminability of distinctive features in normal and whispered speech, J. Acoust. Soc. Am. 90: 2327 (1991).
- 103. S. Wang, A. Sekey, and A. Gersho, An objective measure for predicting subjective quality of speech coders, *IEEE J. Select. Areas Commun.* **10**(5): 819–829 (1992).
- 104. S. W. Wong, An evaluation of 6.4 kbit/s speech codecs for Inmarsat-M system, Proc. ICASSP, 1991, pp. 629–632.
- 105. W. Yang, M. Benbouchta, and R. Yantorno, Performance of the modified bark spectral distortion measure as an objective speech quality measure, *Proc. ICASSP*, 1998, pp. 541–544.
- 106. W. Yang and R. Yantorno, Improvement of MBSD by scaling noise masking threshold and correlation analysis with MOS difference instead of MOS, *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 673–676.
- S. Yeldener, A 4 kbps toll quality harmonic excitation linear predictive speech coder, *Proc. ICASSP*, 1999, pp. 481–484.

SPEECH PERCEPTION

Hannes Müsch GN ReSound Corporation Redwood City, California Søren Buus Northeastern University Boston, Massachusetts

Speech is probably one of the oldest methods of communication. It is produced by an articulatory system constituting the respiratory tract, the vocal cords, the mouth, nasal passages, tongue, and lips. Precisely coordinated actions of all these parts produce a highly complex signal that encodes messages in a very robust manner, which enables the receiver to understand the message even if it is severely degraded by noise or distortion. What exactly makes speech intelligible? A definite answer to this question has yet to come, but a large amount of research has revealed a multitude of factors that are important for speech intelligibility. This article seeks to elucidate some of these factors, especially those that are important for the design of communication systems. A brief description of the speech signal is followed by a brief discussion of some basic properties of speech perception and an overview of some theories that have been proposed to account for them. These theories may be thought to provide a microscopic account of speech perception in

2360 SPEECH PERCEPTION

that they typically seek to explain how listeners map the acoustic properties of the speech signal onto an internal linguistic representation of the message. Other theories take a macroscopic approach. They are not concerned with the specific errors that a listener may make. Rather, they attempt to predict the overall percentage of the speech that will be understood when the speech signal is degraded by noise, distortion, reverberation, and other artifacts that typically are introduced by communication systems and room acoustics. Although these theories reveal relatively little about the processes involved in speech perception, a large portion of this article is devoted to them because they are useful for predicting how the intelligibility of speech is affected by the properties of a communication system.

1. THE SPEECH SIGNAL

Speech is usually considered to consist of a string of individual speech sounds — phonemes — that combine into words, phrases, and sentences. The phonemes correspond to the individual consonants and vowels in a word. At a finer level of analysis, the phonemes are composed of a combination of phonetic features, each of which corresponds to specifics of its production and/or the acoustic properties that signal it. The phonetic features fall into three classes: voicing, place of articulation, and manner of articulation. *Voicing* refers to whether the vocal cords vibrate when the phoneme is produced. Voiced phonemes include all vowels and consonants such as /b/ (as in bat), /z/ (as in zip); in contrast, consonants such as /p/ (as in pat) and /s/ (as in sat) are voiceless. Place of articulation refers to the position of the primary constriction of the vocal tract. Examples are bilabial (e.g., /m/ as in mat), alveolar (e.g., /n/ as in net), and velar (e.g., /g/ as in get). Manner of articulation includes features such as nasal(e.g., /m/ and /n/, which are produced by lowering the soft palate to open the passage to the nasal cavity), stop (e.g., /b/, /p/, and /d/, which are produced by momentarily blocking the vocal tract such that pressure builds up to produce a brief burst of noise when the closure is released), and fricative (e.g., /s/ and /f/, which are produced by turbulent air flow through a constriction of the vocal tract). Whether phonemes, phonetic features, or other units of speech constitute basic units in speech perception remains an open question.

1.1. Temporal Representation

Figure 1 shows the acoustic waveform of a male talker saying the phrase "the happy prince." On the timescale of Fig. 1a, the envelope of the signal is the most prominent feature. Bursts of acoustic energy can be seen clearly. These bursts typically are 100–200 ms apart and often coincide with the syllables of the speech. Note that word boundaries are not clearly marked. There is no pause between "the" and "happy," yet there is a pause-like gap within "happy." The complex envelope of the speech signal encodes considerable information about the message.



Figure 1. Sound pressure-time function of the phrase "the happy prince," spoken by a male talker. The average speech level is 65 dB SPL. (a) The envelope fluctuations are the most prominent feature. (b) Fine structure of the vowel // in "the." (c) Fine structure of the fricative /s/ in "prince."

Enlarging part of the graph reveals the fine structure of the acoustic wave. The fine structure also contains much information about the message. The panel in Fig. 1b shows the trace of the sound pressure during the vowel /a/ in "th<u>e</u>." The sound is not strictly periodic, but there is a high degree of regularity in the signal, which gives rise to a well-defined pitch. The repetition rate of the pattern is called the *pitch period*. The pitch helps convey meaning by prosody (e.g., the pitch rises at the end of a question) and specifying the identity of the speaker, but it is not important for the identity of the vowel.

In contrast, the hissing sound /s/ at the end of "prince" has a noise-like fine structure. Because it is a voiceless consonant, the vocal cords do not vibrate during its production. Rather, the sound is generated by the turbulence that results when air is pressed through a constriction in the vocal tract. The acoustic energy of voiceless consonants typically is much lower than that of vowels.

The speech envelope, which is so obvious in Fig. 1, is obliterated when a masking noise is added to the speech. At negative signal-to-noise ratios (SNRs), the envelope of the speech-and-noise composite is almost flat and the time course of the signal looks almost like that of a noise. Nevertheless, most people have no problems communicating at SNRs somewhat below 0 dB. (As will become evident later, the SNR necessary to just understand speech depends on the spectral shape and temporal characteristics of the noise.)

1.2. Spectral Representation

Many aspects of the speech signal are best described in the frequency domain. Spectra of vowels, for example, have prominent peaks, which represent the resonances of the acoustic filter formed by the vocal tract. They are called formants, and their interrelationships are important in defining the identity of vowels and some features of consonants. Thus, short-term spectra yield a highly useful characterization of speech. They, too, encode much information about the message contained in the speech and are used as a foundation for most automatic speechrecognition schemes. The changing spectral content of the speech signal is often displayed in a spectrogram, which is a power spectrum-time plot. Figure 2 shows the spectrogram of the utterance whose time trace is shown in Fig. 1. The timescale on the abscissa is the same in both figures. Frequency is shown on the ordinate. The signal power at any frequency and point in time is represented by the darkness of the plot. Dark areas correspond to high power and light areas correspond to low power.

The energy fluctuations that were seen in Fig. 1 are also apparent in this presentation. The high-energy periods produced by vocal-cord vibration are easily seen as dark patches, which are separated by light low-energy periods that occur during periods of consonantal noise and during speech pauses. For most speech sounds, the energy is concentrated in the frequency region below 5000 Hz. An exception is the /s/ at the end of "prince," which has most of its energy at high frequencies.

The formants are identified by the dark traces below 3000 Hz. They are visible primarily during the voiced



Figure 2. Spectrogram of the signal shown in Fig. 1. Time is plotted along the abscissa and frequency along the ordinate. The instantaneous signal power is shown by the darkness of the plot. Dark areas represent high power and light areas represent low power.

excitation of the vowels. The formants of a steady-state vowel have constant frequencies, which produce horizontal formant trajectories. The formant frequencies reflect the shape of the vocal tract, which the talker adjusts to produce the desired vowel. Ideally, it should attain a specific shape to produce a given vowel. During natural speech, however, the articulators must move quite rapidly to produce the desired sequence of phonemes and they never reach the ideal positions. Rather, they constantly move from one (unreached) target to the next. As a result, the formants of natural speech show a pattern of rising and falling trajectories that reflect the articulator movement. Because they reflect the movement from one vocal tract shape to the next, the formant trajectories encode not only the vowel identity but also the identity of the preceding and succeeding consonant — at least to some degree. This interaction is called *coarticulation*. When a masking noise masks the low-energy consonants but not the high-energy vowels, a large number of the consonants can still be identified correctly. This is possible because a part of the consonant is encoded in the format transitions that result from coarticulation. Coarticulation also is at the heart of one of the most vexing problems in speech recognition by humans or machines: the lack of a oneto-one correspondence between the acoustic properties of the speech signal and the speech units [1]; (see, however, Ref. 2). For example, the acoustic properties of a speech segment that encodes the phoneme /b/ vary greatly depending on the context (i.e., other phonemes) that surrounds it, as well as on the speaker and speaking rate.

2. SPEECH WITH REDUCED ACOUSTIC INFORMATION

The speech signal encodes the message in the envelope, in the fine structure, and in the short-term spectrum to varying degrees. This produces a redundancy that allows speech to retain a high degree of intelligibility, even if some of these aspects are altered or removed by noise, distortion, or signal processing. For example, Shannon et al. [3] modulated bands of noise with the envelope of the speech signal in these bands. As a result of this operation, the temporal fine structure of the signal was lost completely and the spectral resolution was greatly reduced, but intelligibility remained high. Even when all spectral information was removed—that is, when the broadband envelope of speech was imprinted on a broadband noise—the speech was still recognizable. In this latter condition, the temporal envelope was the only cue available to the listener.

The fact that speech remains intelligible after the entire fine structure has been removed does not mean that only a small fraction of the message is encoded in the fine structure. On the contrary, speech can be highly intelligible when the listeners' only cue is the fine structure. To demonstrate this, Licklider and Pollack [4] passed speech through an infinite peak clipper, so that the processed signal had a fixed positive value whenever the input signal was positive and a fixed negative value whenever the input was negative. This binary code flattens the envelope completely and broadens the spectrum. The processed signal contains information only about the zero crossings of the original signal. Nevertheless, speech processed in this way is intelligible enough to let conversation take place with little difficulty. Even if it sounds very distorted, its quality is high enough to reveal readily the identity of the speaker.

Yet another minimalist representation of speech is sine-wave speech. In this modification, the speech is synthesized by tracing the dominant formants with variable-frequency tone generators. The speech thus synthesized is also highly intelligible, even when only a small number of generators are used [5].

These examples make it clear that speech is very robust. Both spectral and temporal cues carry the message. Speech cannot be defined completely in the spectral domain, nor can it be described only in the temporal domain. The large redundancy allows communication in very unfavorable listening conditions. At the same time, it makes speech a difficult subject to study.

3. SOME BASICS OF SPEECH PERCEPTION

One hallmark of speech perception is that it is categorical. If one constructs a set of speech sounds that gradually changes the acoustic properties from those corresponding to one phoneme (e.g., /d/) to those corresponding to another (e.g., /g/), listeners will report hearing /d/ until the stimuli reach a boundary at which point the perception changes rapidly to become /g/. Discrimination between two members of the same phonemic category usually is difficult, whereas discrimination between members of different phonemic categories is easy [6]. However, listeners can discriminate among stimuli within a single category under ideal conditions [7] and some members of a stimulus set may be shown to be better exemplars than others [8]. Thus, the categorization of speech does not eliminate information about differences between tokens of the same categories. In fact, more recent research indicates that the phonetic categories have a complex internal structure [9], and much current research aims to discover the internal organization of phonetic categories [10,11].

Another hallmark of speech perception is its sensitivity to context. As discussed above, coarticulation makes the acoustic representation of a particular phoneme depend on the context in which it occurs. However, contextual effects are not limited to coarticulation. Variables such as speaking rate and lexical context also may change the perception produced by a given set of acoustical properties. For example, a particular speech sound may be heard as /ba/ if it occurs in a context of slow speech, but as /wa/ if it occurs in a context of rapid speech [12]. Complete theories of speech perception ultimately must explain such context effects, as well as how listeners deal with the lack of invariance between phonemes (or features) and their acoustic representations.

4. THEORIES OF SPEECH PERCEPTION

Various theories have been proposed to account for speech perception at differing levels of detail. Some take a macroscopic approach and seek to predict how the overall percentage of recognition depends on the long-term average properties of speech, distortion, and background noise. This class of theories encompasses the classic articulation index (AI) and derivative theories such as the speech intelligibility index (SII) and speech transmission index (STI). These theories do not attempt to predict how recognition errors depend on the details of the speech signal, but aim to provide a tool that can predict the intelligibility of speech transmitted through a communication system. Because these theories have proved to be very useful engineering tools, they are discussed at some length below. Other theories take a microscopic approach and seek to explain how the detailed properties of the speech signal determine its mapping into linguistic units. This class of theories encompasses a number of very different approaches to various parts of this difficult problem. Some seek to explain speech perception as a result of the general properties of signal processing by the auditory system, sometimes combined with cognitive processing. Others hold that speech recognition is mediated by mapping the acoustic properties of the speech to the processes necessary to produce it. They imply that the recognition of speech is specific to humans. Whether extraction of the phonetic units occurs directly by auditory processes or through reference to articulatory processes, subsequent processing also is important. Thus, other models seek to incorporate the interplay between lexical processes and lower-level phonetic processes.

Auditory theories of speech recognition hold that the auditory processes responsible for perception of any sound generally define the relation between acoustic properties and the categorical perception of phonemes [13], although subsequent lexical, syntactical, and grammatical processing may modify the perception. Specific theories within this class range from auditory processes being responsible for mapping the acoustic properties into features that feed into linguistic processes [14] to ones that claim that a detailed analysis of auditory processing can explain how the categorical boundaries between phonemes depend on context and provide the invariance that appears to be lacking in the acoustical properties of the speech [2]. Holt and co-workers have suggested that many context effects can be explained as a spectral contrast enhancement mediated, in part, by neural adaptation that occurs throughout the auditory pathway [15,16].

Theories that rely on listeners' knowledge of speech production to explain speech perception hold that listeners achieve a mapping between the context-dependent acoustic properties of speech and the corresponding phonemes because they infer how the speech was produced. The most prominent of these theories is the motor theory of speech perception [17]. This theory holds that speech perception is accomplished, at least in part, by a specialized phonetic processor that maps the acoustic signal into the articulatory gestures that underlie the production of it.

Theories on the processing that occurs subsequent to the extraction of phonetic units often hold that speech perception reflects an intimate interaction between lowlevel auditory and phonetic processes and higher-level lexical processes (for a review, see the article by Protopapas [18]). Some employ top-down processing, in which excitatory and inhibitory connections let nodes corresponding to words modify the response properties of nodes corresponding to individual phonemes [19]. Others employ bottom-up processing, in which lower-level nodes of a network activate higher-level nodes in a manner that allows sequences of short-term spectra to map into words [20].

5. THE ARTICULATION INDEX

Although speech intelligibility is determined by many factors, two parameters appear to be of paramount importance: the bandwidth and the audibility of the signal. Both parameters are affected by the transfer function of a transmission system, and it is important for the communication engineer to understand how they affect intelligibility. A group of models, collectively known as articulation index (AI) models, attempt to predict intelligibility from the transfer function of a transmission system. A team led by Harvey Fletcher at Bell Labs laid the groundwork for the development of the Articulation Index in the early 1900s (see Allen [21] for a historical perspective). The complete model was published in 1950 by Fletcher and Galt [22], but is rarely used because it is very complex [23]. A simpler version was published earlier [24]. On the basis of French and Steinberg's model, Kryter [25] devised a model that was easy to use. With minor modifications, Kryter's model was later accepted as ANSI standard S3.5 [26]. The newest member of the group of Articulation-Index models is the speech intelligibility index (SII) [27].

Articulation index models are macroscopic models of speech recognition. They predict the *average* intelligibility achievable with a linear or nearly linear transmission system, where the average is taken across many talkers and speech materials. The model predictions are based on the statistics of the speech signal. Therefore, AI models do not predict the intelligibility of a short speech segment, nor do they account for the intelligibility of particular phonemes and the patterns of confusions among phonemes that occur in the perception of partly intelligible speech [28].

The basic assumption underlying AI theory is that speech intelligibility is determined by the audibility of the speech signal and that different frequency ranges make unequal contributions to the intelligibility. The AI model provides a method to calculate an articulation index. The AI is highly correlated with speech intelligibility and is a descriptor of the intelligibility that can be achieved with the transmission system for which it was calculated. If two people, one meter apart and facing each other, converse in a quiet room that is free of reverberation, intelligibility can reasonably be expected to be optimal. For such a condition, the AI is unity. In this optimal listening condition, all relevant speech cues are completely audible. If, on the other hand, the listener cannot hear the talker at all, there will be no intelligibility. For such a condition, the AI is zero. In general, the AI is a scalar between zero and unity.

Every listening condition has associated with it a single value of AI. Different speech materials yield the same AI-as long as they are equally audible. The AI does not reflect the fact that the intelligibility differs among syllables, words, and various types of sentences presented through a given transmission system. Such differences occur mainly because these materials differ in terms of how much information they provide by context. To account for the effects of context on the speech test score, AI models use different transformation functions between the AI and the predicted test score. These transformations are independent of the speech transmission system and specific to the test material and the scoring procedure. Typical transformations are shown in Fig. 8 (later in this article). Although equivalent to the AI in many respects, the SII (the result of the speech intelligibility index model [27]) differs from the AI by assuming that the SII is not exclusively determined by audibility but to a small degree also by the type of the speech material used [29].

The AI is calculated as an importance-weighted sum of a quantity W_i that is related to the audibility of the speech signal in a set of spectral bands that span the range of audiofrequencies:

$$AI = \sum_{i} I_i W_i \tag{1a}$$

where I_i is the importance weight of the *i*th band. The product of I_i and W_i is the AI in the *i*th band, AI_i. Accordingly, Eq. (1a) can be rewritten as

$$\mathbf{AI} = \sum_{i} \mathbf{AI}_{i} \tag{1b}$$

5.1. Audibility

The quantity W_i is a fraction less than or equal to unity that indicates how many of the cues carried in the *i*th band are actually available to the listener. It is related to the audibility of the speech signal. Audibility is expressed in terms of the *sensation level*, which is the number of decibels that a sound is above its threshold. The sensation level of a band of speech can be calculated as the difference between the critical-band level of the speech and the level of a just-audible tone centered in that band. The critical band is a measure of the bandwidth of frequency analysis in the auditory system. It can be defined as "that bandwidth at which subjective responses [to sound] rather abruptly change" [30] and represents a frequency range across which the acoustic energy is integrated by the auditory system. The critical band is about 100 Hz wide for center frequencies below 500 Hz and 15-20% of the center frequency at higher frequencies [30–32]. [It should be noted that modern measurements of auditory filter characteristics yield an *effective rectangular bandwidth* that is somewhat narrower than the critical bandwidth, especially at low frequencies [33].]

In quiet, the level of a just-audible tone depends only on the listener's sensitivity to sound. The lowest level at which the tone can be heard is called the *absolute threshold*. Threshold levels vary with frequency and among listeners, but established norms exist for normal threshold levels [34,35]. In individuals with hearing loss, thresholds are elevated—that is, the tone levels at threshold are higher than in normal-hearing persons. Therefore, audibility is decreased in listeners with hearing loss. In agreement with the reduced intelligibility of unamplified speech for listeners with hearing loss, the elevated thresholds produce lower AI values.

Interfering sounds (i.e., maskers) also can reduce the audibility of speech. If the masker is above threshold, it may produce masking and make a tone at absolute threshold inaudible. The level of the test tone that can just be detected in the presence of the masker is called the masked threshold. The amount of masking, which is the difference between the masked threshold and the absolute threshold, depends on the level and spectral shape of the masker and the frequency of the tone (for review, see Buus' article [33]). For typical maskers with spectra that change relatively slowly with frequency, the amount of masking is determined primarily by the sound level of the masker measured within a critical band centered on the tone. For such maskers, masking is linear. Whenever the masker level is increased by, for example, 10 dB, the masked threshold increases by 10 dB. For listening situations with masking sounds, the sensation level of the speech is determined as the difference between the critical-band level of the speech and the masked or absolute threshold for tones at the center of the band, whichever is higher.

The relation between band sensation level and W_i differs somewhat across AI models [36]. In their simplest form, AI models assume a linear relation between sensation level and W_i [25–27] over a 30-dB range of sensation levels (see Fig. 3). Within this range W_i increases as the sensation level increases. In the models of French and Steinberg [24] and Fletcher and Galt [22], the relation is nonlinear and the range of sensation levels over which W_i changes is larger than 30 dB.

Because W_i and band AI_i are proportional [see Eq. (1)], the AI_i also increases with sensation level. The level dependence of the AI can be understood by noting that speech is an amplitude-modulated signal. French and Steinberg [24] analyzed data of Dunn and White [37], who measured the level distribution of 125-ms speech segments in a number of frequency bands. The duration of these segments corresponds roughly to the average duration of syllables. When plotting the percentage of



Figure 3. Relation between the audibility-related quantity W_i and the number of decibels that the average speech level is above threshold [25,26]. This function reflects the distribution of the short-term speech level in 125-ms intervals. If the long-term average speech level is 12 dB below the listener's threshold, speech peaks will exceed the listener's threshold and W_i becomes larger than zero.

intervals that exceed a certain level as a function of the difference between that level and the long-term average speech level, French and Steinberg found a relationship that is well described by a linear function. This relationship, together with the masked or absolute threshold, allows one to derive the percentage of syllables that are intense enough to be heard by the listener at any given speech level. The difference between speech level and threshold level is a measure of the proportion of syllables that the listener hears. When this proportion increases, W_i — and the AI — in the band also increase. According to Kryter [25], approximately 1% of the 125-ms intervals of speech is audible when the long-term average speech level is 12 dB below detection threshold. At this point W_i is just above zero. From there, W_i increases linearly with level as a larger proportion of the syllables becomes audible. Once the long-term average speech level is 18 dB above threshold, even the weakest percentile of 125-ms intervals exceeds threshold. At this point, W_i is unity and does not increase further with additional level increases.

5.2. Independence Assumption

As already stated, the AI and the intelligibility score are related by a nonlinear transformation function that is different for different test materials. With one exception, these transformations are derived empirically. Only the transformation between AI and the proportion of phonemes correctly identified in a nonsense-syllable recognition task, also known as the *articulation*, *s*, is determined by the structure of the AI model. Articulation and AI are related by

$$\mathbf{AI} = -k \cdot \log_{10}(1-s) \tag{2}$$

where k is a scale factor that is chosen such that AI is unity in the most favorable listening condition. Equation (2) results from the fundamental assumption that speech
bands contribute independently to articulation. This assumption underlies most AI models (see, however, Ref. 38). Combining Eqs. (1) and (2) yields

$$\log_{10}(1-s) = \sum_{i} \log_{10}(1-s_i)$$
(3)

which is equivalent to

$$(1-s) = \prod_{i} (1-s_i) \tag{4}$$

where s is the articulation in the broadband condition and s_i is the articulation achieved when listening to the individual bands. The terms (1 - s) and $(1 - s_i)$ are the associated error probabilities. Equation (4) states that the probability of identifying a phoneme in a nonsense syllable incorrectly is equal to the product of the error probabilities when decisions are based on the information available in the individual bands. Whenever a joint probability equals the product of the conditional probabilities, the events are said to be independent. Therefore, Eqs. (3) and (4) reflect the AI model assumption that speech bands contribute independently to the intelligibility of phonemes in nonsense syllables.

5.3. Band Importance

The importance of various frequency bands for speech understanding varies with frequency. This can be seen in Fig. 4, which shows the proportion of correctly understood phonemes in highpass- and lowpass-filtered nonsense syllables as a function of the filters' cutoff frequencies. The speech is at a level that ensures that audibility in the passband is always unity. For the purpose of speech intelligibility predictions, a lowpass filter with a cutoff frequency of 8000 Hz is equivalent to an allpass, because the speech signal carries very little energy in the bands above 8000 Hz. By definition, the AI of a system with a flat transfer function that provides optimal gain is unity. The data in Fig. 4 show that articulation is almost perfect (s = 0.985) for this ideal transmission system. Likewise, when the cutoff frequency of a highpass filter is 100 Hz, the same high performance is observed because the acoustic information in bands below 100 Hz is irrelevant for speech intelligibility.

Performance decreases when the low-frequency components of the speech are gradually removed by increasing the cutoff frequency of the highpass filter. Initially, the performance decrease is only slight. Raising the cutoff frequency from 100 to 1000 Hz affects articulation only marginally. This suggests that the speech bands between 100 and 1000 Hz make only small contributions to intelligibility. When the cutoff frequency is raised further, articulation decreases sharply. Removing the frequency range between 1000 and 4000 Hz causes articulation to drop from almost perfect to just above 0.5. Apparently, this frequency band carries a significant portion of the speech cues.

Similarly, performance decreases when the highfrequency speech components are removed by decreasing the cutoff frequency of the lowpass filter. Again, performance declines very gradually at first, but the decrease becomes quite rapid once the cutoff frequency is below 5000 Hz. The frequency at which the two curves intersect divides the speech spectrum into two equally important bands. This frequency, called the *crossover frequency*, is approximately 1700 Hz. Because the AI is unity for a broadband condition in which the speech is clearly audible across the entire frequency range, it follows that the AI for each of the two equally important bands must be 0.5.

Equation (1a) states that the AI is an importanceweighted sum of W_i , which is related to the audibility in the bands. The *importance density function* describes how the inherent potential for intelligibility is distributed across frequency. It can be derived from the relation between the filter cutoff frequency and the associated



Figure 4. Crossover functions. The proportion of correctly identified phonemes in a nonsense-syllable context is plotted as a function of the cutoff frequencies of ideal highpass and low-pass filters. Filled dots and crosses represent highpass filters; open dots and plusses, lowpass filters. (Reprinted from Fletcher and Galt [22], with permission of the publisher.)

articulation. Using Eq. (2), the articulation scores are transformed into the corresponding AI values. This results in a relation between filter cutoff frequency and AI. Thus with the help of Eq. (2), the data in Fig. 4 can be transformed into a cumulative importance function. Simple differentiation yields the importance-density function. The band-importance weights, I_i , are derived from the importance-density function by integrating over the width of the band. Figure 5 shows the importance weights for critical-band wide bands of nonsense syllables [27].

The various AI models differ in the details of the calculation [39]. For example, different models use different importance functions. The exact shape of the importance function seems to depend on the speech material used [29,40]. Accordingly, the SII specifies several speech-material-specific importance functions. Differences also exist in the degree of sophistication with which the sensation level of the speech is predicted from parameters of the transmission system and the spectrum of the masking noise. The representation of the speech spectrum and the masker spectrum in the inner ear is "smeared" in both the spectral and temporal domains. Not only can the masker mask the speech, but the speech signal also acts as a masker upon itself. The amounts of noise-masking and self-masking depend on the speech level and on the transfer function of the transmission system and can be predicted quite accurately. The model by Fletcher and Galt [22] and the SII [27] exhibit the most sophistication in modeling both forms of masking. The simple AI of ANSI [26] does not model self-masking of speech and its prediction accuracy is compromised if the transmission system filters speech so that high-level portions of the spectrum are adjacent to low-level portions.

The AI model discussed so far predicts that when speech exceeds an optimal level, further level increases do not result in increased intelligibility. However, intelligibility may decrease at excessive sound levels [22,41]. This effect is known as *rollover*. The AI by Fletcher and Galt [22] and the SII [27] model this effect.



Figure 5. The importance function for nonsense syllables as defined by the speech intelligibility index (SII). Every data point represents the integral of the importance density function over a frequency range equal to one critical bandwidth centered at the frequency of the symbol. This importance function is in good agreement with the data in Fig. 4.



Figure 6. Speech intelligibility index (SII) of undistorted speech as a function of speech level. The predictions are for speech reception in quiet by normal-hearing listeners.



Figure 7. The sound pressure level (SPL) at the detection threshold (bold line, ISO 226) is plotted together with the long-term average level of speech in critical bands (solid lines labeled with the overall speech level). The dashed lines indicate the dynamic range of the speech (from 12 dB above to 18 dB below the long-term average).

Figure 6 shows the SII of unfiltered speech as a function of the overall speech level. At very low speech levels, the signal is inaudible and the SII is zero. As the speech level increases, spectral components in the critical band centered at 500 Hz are the first to exceed threshold. This can be seen from Fig. 7, which shows the long-term average critical-band level of speech with an overall level of 4 dB SPL (lower solid line) in relation to the threshold of normally hearing listeners (bold line [35]). The dashed line 12 dB above the long-term average speech level marks Kryter's [25] estimate of the top of the dynamic range of the speech signal.

The rate at which the SII increases with speech level depends on the sum of the importance weights in the bands whose audibility is affected by the level change. At low levels, the SII increases only slowly with level because only the bands near 500 Hz contribute. At higher levels, all bands are audible and a 3-dB change in speech level changes the SII by 0.1. This rate is equal to the rate at which W_i increases with sensation level (see Fig. 3) and is the steepest slope possible.

Once the speech level reaches 59 dB SPL, all speech bands are completely audible. Figure 6 shows that at this level the SII is unity, indicating that intelligibility is optimal. This is also seen in Fig. 7, where the critical-band level of 59-dB-SPL speech is shown by the upper solid line. The dashed line 18 dB below represents Kryter's [25] estimate of the lower bound of the dynamic range. In frequency bands that carry speech cues, this lower bound is above threshold, indicating that all speech cues are available to the listener. Further increasing the speech level does not result in increased intelligibility. On the contrary, when the speech level increases above 68 dB SPL, the SII decreases.

5.4. Transformation Between Articulation Index and Intelligibility

The articulation index describes the extent to which the sensation generated by the stimulus itself affects intelligibility. However, speech recognition depends not only on the sensory input but also on the context in which it occurs. Semantic and syntactic context, lexical constraints, and the size of the test set strongly affect intelligibility. None of these factors are reflected in the AI (some are reflected in the SII, however). Instead, they are accounted for by a set of transformation functions between AI and the predicted intelligibility score. Because these transformations depend strongly on the contextual constraints in the speech material and on the scoring procedure, their form varies significantly. One set of transformation functions that can serve as a reference is reproduced in Fig. 8. However, many users of AI models prefer to derive transformation functions specifically for their particular speech materials. In fact, the SII [27] states that users must define the appropriate transformations for the particular speech material and scoring method being used. Whatever transformation is used, those shown in Fig. 8 are reasonably representative. They indicate that for highly redundant materials, such as everyday speech, an AI of about 0.2 is needed to obtain a marginal intelligibility of 75%.

It has been shown that words in meaningful sentences are more easily understood than words in isolation or words preceded by a carrier phrase. Lexical constraints cause the intelligibility of phonemes in meaningful words to be higher than the intelligibility of phonemes in nonsense syllables. In general, intelligibility increases as the predictability of the speech sounds increases.

By applying optimal decision theory to speech recognition, Pollack [42] has shown that the increased recognition performance for items with high *a priori* probability of occurrence cannot be explained by more successful guessing. When he calculated the receiver operating characteristics (ROCs) that describe the speech-recognition performance for speech tokens with different *a priori* probabilities of occurrence, he found that the observers were more sensitive to speech tokens with a high *a priori* probability of occurrence than to those with a low *a priori* probability. The better performance for items with high priori probability was not the result of a response



Figure 8. Transformation functions between the AI and the predicted intelligibility of various speech materials. The more contextual constraints are placed on the signal, the higher the intelligibility. (Reprinted from Kryter [25], with permission of the publisher.)

bias. Rather, it appears as if context provides an additional source of information that increases the listener's sensitivity to speech recognition. It has been shown that a signal-detection model may account for such effects by changing the amount of variance contributed to the decision variable by a single, central noise source [43].

An entirely different method of accounting for context effects assumes that contextual information is statistically independent from the sensory information used to recognize the target without context [44]. Expanding on the idea of multiplicative error probabilities and assuming that the context is extracted from speech presented in the same listening condition as the sensory input, this model derives a simple relation between the probability, p_c , of correctly recognizing items in the presence of context and the probability, p_i , of recognizing the same items without context:

$$p_c = 1 - (1 - p_i)^a \tag{7}$$

When this equation was applied to several sets of data, a was nearly constant over a wide range of values of p_i , lending support to the assumption of an independent contribution of the context [44].

5.5. The Speech Transmission Index

The AI model definition of audibility implicitly assumes that the temporal structure of the speech signal is undisturbed and that the short-term level of the masking noise is unrelated to the short-term level of the speech signal. AI models break down when these requirements are not met. To circumvent this problem, the speech transmission index (STI) [45] converts temporal distortions of the speech signal into an equivalent reduction in the audibility of the speech. This approach extends the AI concept to include temporal distortions such as reverberation and nonlinearities (e.g., a level-compression circuit), which can reduce the modulation depth of an amplitude-modulated signal such as speech. The reduction of modulation depth caused by temporal distortion usually varies across modulation frequencies. For example, reverberation tends to reduce fast modulations more than slow modulations. Therefore, the STI model determines the effective SNR in every frequency band for a set of modulation frequencies ranging from 0.63 to 12.5 Hz, which are typical for speech. Any reduction in modulation depth is transformed into an equivalent noise level, which yields a transmission index for each modulation frequency. The average of the transmission indices at a given spectral frequency band yields an overall modulation transmission index for the band, which is largely equivalent to the AI contribution by the band. Thus, the STI is very similar to the AI models, but its reliance on modulation transfer allows it to account for situations that the AI models do not handle readily.

6. SUMMARY

Speech is a signal in which a message is encoded. The coding is not yet fully understood, but it is known that it is very robust, because speech remains intelligible even after severe signal distortion. Speech perception is categorical. Theories of speech perception attempt to explain the speech-perception process. Examples of such theories are auditory theories [e.g., 2] and the motor theory of speech perception [17]. Although the process of speech recognition and comprehension is not yet fully understood, models exist that predict human speechrecognition performance without attempting to explain the speech perception process. These models are statistical in nature and predict only the average intelligibility. Examples of such models are the speech intelligibility index (SII), the speech transmission index (STI), and the articulation index (AI).

BIOGRAPHIES

Hannes Müsch received the Diplom Ingenieur degree in acoustics and electrical engineering in 1993 from the Technische Universität Dresden, Germany, and the MSEE and Ph.D. degrees in electrical engineering from Northeastern University, Boston, Massachusetts, in 1997 and 2000, respectively. Dr. Müsch was a consultant in industrial and building acoustics at Müller-BBM, Germany, and a research scientist at GN ReSound's Core Technology Center in Redwood City, California, where he worked on the development of signal processing algorithms for hearing aids. In 2001, he joined Sound ID, Palo Alto, California, where he is the director of signal processing. Dr. Müsch's research interests are psychoacoustics, models for the prediction of speech recognition performance, and the effects of hearing loss on audition. His work is focused on developing signal processing algorithms to improve the listening experience for the hearing impaired.

Søren Buus is the director of the Communications and Digital Signal Processing Center and professor of electrical and computer engineering at Northeastern University Boston, Massachusetts. He graduated with an M.S. degree in electrical engineering and acoustics from the Technical University of Denmark in 1976 and a Ph.D. in experimental psychology from Northeastern University in 1980. His primary research interests are in basic psychoacoustics. In particular, he has published extensively on intensity coding, detection, loudness, and temporal processes in the auditory system. Before joining the faculty at Northeastern University in 1986, Professor. Buus held research appointments at Northeastern University and Harvard University, Massachusetts. He also has been a guest researcher at the Acoustics and Mechanics Laboratory, CNRS, Marseille, France; Institute of Electroacoustics at the Technical University of Munich, Germany; Department of Environmental Psychology, Faculty of Human Sciences, Osaka University, Japan; and the Acoustics Laboratory, Technical University of Denmark. He is a fellow of the Acoustical Society of America.

BIBLIOGRAPHY

- J. L. Miller, Speech perception, in M. C. Crocker, ed., *Ency-clopedia of Acoustics*, Vol. 4, Wiley, New York, 1997, pp. 1579–1588.
- K. N. Stevens and S. E. Blumstein, The search for invariant acoustic correlates of phonetic features, in P. D. Eimas and J. L. Miller, eds., *Perspectives on the Study of Speech*, Erlbaum, Hillsdale, NJ, 1981, pp. 1–38.
- R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, Speech recognition with primarily temporal cues, *Science* 270: 303-304 (1995).
- J. C. R. Licklider and I. Pollack, Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech, J. Acoust. Soc. Am. 20: 42-51 (1948).
- R. E. Remez, P. E Rubin, D. B. Pisoni, and T. D. Carrell, Speech perception without traditional speech cues, *Science* 212: 947–950 (1981).
- A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, The discrimination of speech sounds within and across phoneme boundaries, J. Exp. Psychol. 54: 358-368 (1957).
- A. E. Carney, G. P. Widin, and N. F. Viemeister, Noncategorical perception of stop consonants differing in VOT, J. Acoust. Soc. Am. 62: 961–970 (1977).
- A. G. Samuel, Phonemic restoration: Insights from a new methodology, J. Exp. Psychol.: Gen. 110: 474-494 (1982).
- P. K. Kuhl, Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not, *Percept. Psychophysiol.* 50: 93-107 (1991).
- J. L. Miller, On the internal structure of phonetic categories: A progress report, *Cognition* 50: 271–285 (1994).
- 11. P. Iverson and P. K. Kuhl, Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from

a common mechanism? *Percept. Psychophysiol.* **62**: 874–876 (2000).

- J. L. Miller and A. M. Liberman, Some effects of lateroccurring information on the perception of stop consonant and semivowel, *Percept. Psychophysiol.* 25: 457-465 (1979).
- J. Coleman, Cognitive reality and the phonological lexicon: A review, J. Neuroling. 11: 295–320 (1998).
- D. B. Pisoni, Auditory and phonetic memory codes in the discrimination of consonants and vowels, *Percept. Psychophysiol.* 13: 253-260 (1973).
- L. L. Holt, A. J. Lotto, and K. R. Kluender, Neighboring spectral content influences vowel identification, J. Acoust. Soc. Am. 108: 710-722 (2000).
- L. L. Holt and K. R. Kluender, General auditory processes contribute to perceptual accommodation of coarticulation, *Phonetica* 57: 170-180 (2000).
- A. M. Liberman and I. G. Mattingly, The motor theory of speech perception, revised, *Cognition* 21: 1-36 (1985).
- A. Protopapas, Connectionist modeling of speech perception, Psychol. Bull. 125: 410–436 (1999).
- J. L. McCelland and J. L. Elman, The TRACE model of speech perception, Cogn. Psychol. 18: 1–86 (1986).
- 20. D. H. Klatt, Speech perception: A model of acoustic-phonetic analysis and lexical access, J. Phonet. 7: 279–312 (1979).
- J. B. Allen, How do humans process and recognize speech? IEEE Trans. Speech Audio Process. 2: 567–577 (1994).
- 22. H. Fletcher and R. Galt, Perception of speech and its relation to telephony, J. Acoust. Soc. Am. 22: 89–151 (1950).
- H. Müsch, Review and Computer implementation of Fletcher and Galt's method of calculating the Articulation Index, *Acoust. Res. Lett. Online* 2: 25–30 (2001).
- 24. N. R. French and J. C. Steinberg, Factors governing the intelligibility of speech, J. Acoust. Soc. Am. **19**: 90–119 (1947).
- 25. K. Kryter, Method for the calculation and use of the Articulation Index, J. Acoust. Soc. Am. 34: 1689-1697 (1962).
- 26. ANSI, S3.5-1969, American National Standard Methods for the Calculation of the Articulation Index, American National Standards Institute, New York, 1969.
- 27. ANSI, S3.5-1997, American National Standard Methods for Calculation of the Speech Intelligibility Index, American National Standards Institute, New York, 1997.
- G. A. Miller and P. E. Nicely, An analysis of perceptual confusion among some English consonants, J. Acoust. Soc. Am. 27: 338-352 (1954).
- 29. C. V. Pavlovic, Factors affecting performance on psychoacoustic speech recognition tasks in the presence of hearing loss, in G. A. Studebaker and I. Hochberg, eds., *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, Needham Heights, MA, 1993.
- B. Scharf, Critical bands, in J. V. Tobias, ed., Foundations of Modern Auditory Theory, Vol. I, Academic Press, New York, 1970, pp. 157–202.
- E. Zwicker, Subdivision of the audible frequency range into critical bands (Frequenzgruppen), J. Acoust. Soc. Am. 33: 248 (1961).
- 32. E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichtenempfänger*, Hirzel-Verlag, Stuttgart, Germany, 1967 (available in Engl. transl. by H. Müsch, S. Buus, and M. Florentine

as *The Ear as a Communication Receiver*, Acoust. Soc. Am., Woodbury, NY, 1999).

- 33. S. Buus, Auditory masking, in M. J. Crocker, ed., Encyclopedia of Acoustics, Vol. 3, Wiley, New York, 1997, pp. 1427–1445.
- 34. ANSI, Specifications for Audiometers, American National Standards Institute, New York, 1989.
- 35. ISO 389-7, Acoustics—Reference Zero for the Calibration of Audiometric Equipment—Part 7: Reference Threshold of Hearing under Free-Field and Diffuse-Field Listening Conditions, International Organization for Standardization, Geneva, 1996.
- C. V. Pavlovic, Derivation of primary parameters and procedures for use in speech intelligibility predictions, J. Acoust. Soc. Am. 82: 413-422 (1987) [erratum: J. Acoust. Soc. Am. 83: 827 (1987)].
- H. K. Dunn and S. D. White, Statistical measurements on conversational speech, J. Acoust. Soc. Am. 11: 278–288 (1940).
- H. Steeneken and T. Houtgast, Mutual dependence of the octave-band weights in predicting speech intelligibility, *Speech Commun.* 28: 109-123 (1999).
- C. V. Pavlovic and G. A. Studebaker, An evaluation of some assumptions underlying the articulation index, J. Acoust. Soc. Am. 74: 1606-1612 (1984).
- 40. G. A. Studebaker and R. L. Sherbecoe, Frequency-importance functions for speech recognition, in G. A. Studebaker and I. Hochberg, eds., *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, Needham Heights, MA, 1993.
- G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney, Monosyllabic word recognition at higherthan-normal speech and noise levels, J. Acoust. Soc. Am. 105: 2431-2444 (1999) [erratum: J. Acoust. Soc. Am. 106: 2111 (1999)].
- I. Pollack, Message probability and message reception, J. Acoust. Soc. Am. 36: 937-945 (1964).
- H. Müsch and S. Buus, Using statistical decision theory to predict speech intelligibility, I Model structure, J. Acoust. Soc. Am. 109: 2896-2909.
- A. Boothroyd and S. Nittrouer, Mathematical treatment of context effects in phoneme and word recognition, Acoust. Soc. Am. 84: 101-114 (1988).
- H. J. M. Steeneken and T. Houtgast, A physical method for measuring speech-transmission quality, J. Acoust. Soc. Am. 67: 318-326 (1980).

SPEECH PROCESSING

DOUGLAS O'SHAUGHNESSY INRS-Telecommunications Montreal, Quebec, Canada

1. INTRODUCTION

When a computer is used to handle the acoustic signal we call speech, sound is captured by a microphone, digitized for storage in the computer, and then manipulated or transformed to serve some useful purpose, such as efficient coding for transmission or analysis for automatic conversion to text. Such processing of speech signals by computer converts the speech pressure wave variations (which actually constitute speech) to other forms of information that are more useful for practical computer applications.

Speech is normally produced by a human speaker, whose brain does the processing needed to generate the speech signal. Similarly, human listeners receive speech through their ears and process it with their brains to understand the spoken message. In this chapter, we examine ways that computers use to analyze speech and to simulate production or perception of speech signals. There are many applications for this speech processing: text-tospeech synthesis (e.g., a machine capable of reading aloud to the visually handicapped; a system able to respond vocally from a textual database), speech recognition (e.g., controlling machines via voice; fast entry into databases), speaker verification (e.g., using one's voice as an identifier instead of fingerprints or passwords), and speech coding (i.e., compressing speech into a compact form for storage or transmission, and subsequent reconstruction of the speech when and where needed). All these uses require converting the original speech into a compact digital form, or vice versa, and related digital signal processing.

Such processing usually involves either analysis or synthesis of speech. Recognition of speech (where the desired result is the text that one usually associates with the speech) or of speakers (where the output is the identity of the source of the speech, or a verification of a claimed identity) requires analysis of the speech, to extract relevant (and usually compact) features that characterize how the speech was produced (e.g., features related to the shape of the vocal tract used to utter the speech). Textto-speech synthesis, on the other hand, tries to convert normal text into an understandable synthetic voice. For speech coding, both the input and output of the process are in the form of speech; thus text is not directly involved, and the objectives are a compression of the data rate (for efficient and secure transmission) and a retention of the naturalness and intelligibility of the original speech, during the analysis and synthesis stages.

2. SPEECH ANALYSIS

For speech analysis, we first convert the speech waves (which, like all analog signals in nature, vary continuously in both time and intensity) into a bitstream, that is, analog-to-digital (A/D) conversion, so that a computer can process the speech signal. The resulting information or "bit rate" (which is of prime interest for efficient operations) is controlled by two factors: the sampling rate (number of evaluations or samples per second) and the sampling precision (number of bits per sample). The sampling rate is directly proportional to the allowed bandwidth of the speech; the Nyquist theorem specifies that the rate be at least twice the highest frequency in the processed signal (otherwise, inevitable "aliasing" distortion appears in the later D/A conversion) [1]. The minimum rate used in practical speech applications is usually 8000 samples/s, thus retaining a range from very low frequencies (theoretically, 0 Hz, although no useful audio data occur below 50 Hz) up to 4 kHz (in practice, an analog lowpass filter usually precedes the A/D converter, and there is a gradual cutoff of frequencies below 4 kHz). Such a low rate is standard for telephone applications, because the switched network only preserves approximately the 300–3200-Hz range. The maximal rate typically employed is 44,100/second (used for audio on compact disks [2–5]), which preserves frequencies beyond the normal hearing range.

Unlike the sampling rate, which follows immediately from the selection of bandwidth range, the choice of bits/sample corresponds to how much distortion is tolerable in a given speech application. To minimize distortion to inaudible levels after A/D conversion, we typically need 12 bits/sample. This gives about 60 dB (decibels) in signal-to-noise ratio, which allows for a typical 30-dB range between strong vowels and weak fricatives in speech. Hence, 12-bit A/D converters are acceptable, although 16-bit systems are more common. In the telephone network, where bandwidth is at a high premium, 8-bit logarithmic (nonuniform) quantization is used on most digital links.

The objective of speech analysis is to transform speech samples into a more useful information representation, for specific objectives other than simply furnishing speech to human ears. The bit rate of basic digital speech (via simple or logarithmic A/D conversion) is typically 64 kbit/s or higher. If we compare that rate to the amount of fundamental information in the signal, we can see a large discrepancy. An average speaking rate is about 12 sounds or phonemes/s, and most languages have an inventory of about 32 phonemes. Thus the phonemic sequence of speech can be sent in approximately 60 bps (bits per second) $(12 \log_2 32)$. This does not include intonational or speaker-specific aspects of the speech, but ignores the fact that phoneme sequences are not random (e.g., Huffman coding can reduce the rate). The overall information rate in speech is perhaps 100 bps. Current speech coders are far from providing transparent coding at such rates, but some complex systems can reduce the rate to about 8 kpbs (at 8000 samples/s) without losing quality (other than limiting bandwidth to 4 kHz).

Typical analysis methods try to extract features that correspond to some well-known aspects of speech production or perception. For example, experiments have shown that speakers can easily control the intensity of speech sounds (and that listeners can easily detect small changes in intensity) [6]. Thus intensity (or a related parameter, energy) is often determined in speech analysis, by simply summing a sequence of (squared) speech samples. Similarly, the positions of resonance peaks in the amplitude spectrum of speech (known as formants, abbreviated F1, F2, ..., in order of increasing frequency) have been correlated with shapes of the vocal tract in speech production; they are also well discerned perceptually (peaks are much more salient than spectral valleys). Analysis techniques typically try to extract compact parameterizations of these spectral peaks, either modeling the underlying resonances of the vocal tract (including resonance bandwidths) or some form of the detail (both coarse and fine) in the amplitude spectrum.

Another feature that is often examined in speech analysis is that of the fundamental frequency of the vocal cords, which vibrate during "voiced" (periodic) sounds such as vowels. The rate (abbreviated F0) is directly controlled in speech production, and the resulting periodicity is easily discriminated by listeners. It is used in tone languages directly for semantic concepts, and cues aspects of stress and syntactic structure in many languages. While F0 (or its inverse, the pitch period — the time between successive closures of the vocal cords) is rarely extracted for speech recognizers, it is often used in low-bit-rate speech coders and must be properly modeled for speech synthesis.

3. SPEECH CODING

The objective of speech coding is to represent speech as compactly as possible (i.e., few bps), while retaining high intelligibility and naturalness, with minimal time delay in inexpensive hardware. There are many practical compromises possible here, ranging from zero-delay, cheap, transparent, high-rate PCM (pulse-code modulation) to systems that trade off naturalness for decreased bit rates and increased complexity [5]. We distinguish higherrate waveform-based coders (usually yielding high-quality speech) from lower-rate parametric coders. Coders in the first class reconstruct speech signals sample-by-sample, representing each speech sample directly with at least a few bits, while the latter class typically discards spectral phase information and processes blocks ("frames") of many samples at a time, which allows the average number of bits/sample to be one or below (e.g., speech coding at 8 kbps).

3.1. Structure in Speech That Coders Exploit

Coders typically try to identify structure in signals, extract it for efficient representation, and leave the remaining (less predictable) signal components either to be ignored (in low-rate systems) or coded using simple techniques (in high-rate systems). There are many sources of structure in speech signals (which, indeed, account for the large difference between typical rates of 64 kbps and an approximately 100-bps theoretical limit). Phonemes typically last 80 ms (e.g., approximately 10 pitch periods), largely due to the increasing difficulty of articulating speech more rapidly (but also to avoid losses in human perception at higher speaking rates). However, identification of each sound is possible from a fraction of each pitch period (at least for speech in noise-free environments). The effective repetition of information in multiple periods helps increase redundancy in human speech communication, which allows reliable communication even in difficult conditions.

The spectra of most speech show a regular structure, which is the product of a periodic excitation (at a rate of F0) and the set of resonances of the vocal tract (which appears as a series of peaks and valleys, averaging about one peak every 1 kHz). Rather than code the speech waveform sample by sample or the spectrum point by point, efficient coders extract from the speech some parameters directly related to the overall amplitude, F0, and the spectral peaks, and use these for transmission.

A simple spectral measure for speech analysis is the zero-crossing rate (ZCR), which provides a basic frequency estimate for the major energy concentration. The ZCR is just the number of times the speech signal crosses the time axis (i.e., changes algebraic sign) in a given time period (e.g., taking an overly simple nonspeech case, a sinusoid of 100 Hz has a ZCR of 200/s). Background acoustic noise often has a steady, broad lowpass spectrum and thus has a ZCR corresponding roughly to a stable frequency in the low range of the signal bandwidth. On the other hand, for weak speech obstruents that are difficult to detect against background noise, the ZCR is either high (corresponding to a high-frequency concentration of energy in fricatives and stop bursts) or very low (if a "voicebar," corresponding to radiation of F0 through the throat, dominates). The ZCR is easy to compute, and can be used for speech endpoint detection (see below).

3.2. Exploiting Simple Structure

Advanced coders examine blocks of data, consisting of sequences of speech samples (which necessarily increases the response time, but facilitates temporal exploitation). The simplest coders (PCM) represent each sample independently. Using a uniform quantifier, however, is inefficient because most speech samples tend to be small (i.e., speech is more often weak than strong), and thus the quantifier levels assigned to large samples are rarely used. Optimal encoding occurs when, on average, all levels are equally used. Since the distribution of speech samples typically resembles a Gamma probability density function (their likelihood decaying roughly exponentially as amplitude increases), a logarithmic compression prior to uniform quantization is useful (e.g., the μ -law or A-law log PCM, common in telephone networks).

Many coders adapt in time, exploiting slowly changing characteristics of the speech and/or transmission channel, such as following the shape or movements of the vocal tract. For example, the step size of the quantifier can be adjusted to follow excursions of the speech signal, integrated over time periods ranging from 1 to 100 ms. When the signal has large energy, using larger step sizes can avoid clipping (which causes nonlinear and severe distortion). When the step size is reduced for low-energy samples, the quantization noise is proportionally reduced without clipping.

3.3. Exploiting Detailed Spectral Structure

Many speech coders use linear prediction methods to estimate a current speech sample based on a linear combination of previous samples [8]. This is useful when the speech spectrum is nonuniform (pure white noise, with independent samples and a flat spectrum, would allow no gain through such prediction). The simplest prediction occurs in delta modulation, where one previous speech sample directly provides the estimate of each current sample; the assumption is that a typical sample changes little from its immediately prior neighbor. This becomes truer if we sample well above the Nyquist rate, as is done in delta modulation, which in turn allows the use of a one-bit quantifier (trading off sampling rate for bit precision) [7].

The power of linear prediction becomes more apparent when the order of the predictor (i.e., the window or number of prior samples examined) is about 10 or so, and when the predictor adapts to movements of the vocal tract (e.g., is updated every 10-30 ms). This occurs in two very popular speech coders, ADPCM (adaptive differential PCM) and LPC (linear predictive coding). They exploit the fact that the major excitation of the vocal tract for voiced speech occurs at the time that the vocal cords close (once per pitch period), which causes a sudden increase in speech amplitude, after which the signal decays exponentially (with a rate inversely proportional to the bandwidth of the strongest resonance — usually F1). Each pitch period approximately corresponds to the impulse response of the vocal tract, and is directly related to its resonances. Since there are about four such formants (in a typical 0-4-kHz bandwidth) and since each formant can be directly characterized by a center frequency and a bandwidth, a predictor order of ten or so is adequate (higher orders give more precision, but offer diminishing returns as computation and bit rate increase).

The 10–16 LPC parameters derived from such an analysis form the basis of many coders, as well as for speech recognizers. These parameters are the multiplier coefficients of a direct-form digital filter, which, when used in a feedforward fashion, converts a speech signal being modeled into a "residual" or "error" signal, which is then suitable for either simple coding (with fewer bits, typically 3-4/sample, in ADPCM) or further parameterization (as in LPC systems).

Using the coefficients in a feedback filter, the system acts as a speech synthesizer, when excited by an impulse train (impulses spaced every 1/F0 samples) or a noise signal (simulating the frication noise generated at a narrow constriction in the vocal tract). This latter, dual excitation is a very simple, but powerful, model of the LPC residual signal, which allows very low bit rates in basic LPC systems [9]. The result is intelligible but synthetic-quality speech (less natural than the "tollquality" speech of the telephone network with log PCM, or with other medium-rate, waveform-coding methods). The LPC model, with its dozen spectral coefficients parameterizing the vocal tract shape (filter) and its three excitation parameters (F0, amplitude, and a single "voicing" bit noting a decision whether the speech is periodic) modeling the residual, often operates around 2.4 kbps (using updates every 20 ms). The LPC coefficients are usually transformed into a more efficient set for coding, such as the reflection coefficients (which are the multipliers in a lattice-form vocal tract filter, and can actually correspond to reflected energy in simple three-dimensional vocal tract models, modeling the two traveling waves of pressure, one going up the vocal tract, the other down). Another popular set is the line spectral frequencies (LSF), which displace the resonance poles in the digital spectral z plane onto the unit circle, which allows more efficient differential coding in one dimension around the circle, rather than the implicit twodimensional coding of the resonances inherent in other LPC forms.

3.4. Exploiting Structure Across Parameters

More efficient coders go beyond simple temporal and spectral structure in speech, to exploit correlations across parameters, both within and across successive frames of speech. Even relatively efficient spectral representations such as the LSFs do not produce orthogonal parameter sets; that is, there remain significant correlations among sets of parameters describing adjacent frames of speech. Shannon's theorem states that it is always more efficient to code a signal in vector form, rather than code the samples or parameters as a succession of independent numbers. Thus, we can group related parameters as a block and represent the set with a single index. For example, to code a speech frame every 20 ms, a set of 10 reflection coefficients might need 50 bits as scalar parameters (about 5 bits each), but need only 10 bits in vector quantization (VQ). In the latter, we chose $1024 \ (= 2^{10})$ representative points in 10-dimensional space (where each dimension corresponds to a parameter). The points are usually chosen in a training phase (i.e., coder development) by examining many minutes of typical speech (ranging over a wide variety of different speakers and phonemes). Each frame provides one point in this space, and the chosen points for the VQ are the centroids of the most densely populated clusters. Since there are only about 1000 different spectral patterns easily discriminable in speech perception (ignoring the effects of F0 and overall amplitude), a 10-bit system is often adequate. Ten-bit VQ represents effectively a practical limit computationally as well, since a search for the optimal point among much more than 1024 possibilities every 20 ms may exceed hardware capacity. VQ basically trades off increased computation during the analysis stage for lower bit rates in later transmission or storage. The most common use today for speech VQ is in CELP (code-excited linear prediction) speech, where short sequences of LPC residual signals are stored in a codebook, to excite an LPC filter (whose vocal tract spectra are themselves represented by another codebook). CELP provides toll-quality speech at low rates, and has become very popular in recent years.

4. FUNDAMENTAL FREQUENCY (F0) ESTIMATION (PITCH DETECTORS)

Both low-rate speech coders and text-to-speech synthesizers require estimation of the F0 of speech signals, as well as the related (and simpler) estimation of the presence or absence of periodicity (i.e., voicing). For many speech signals, it is a relatively simple task to detect periodicity and measure the period. However, despite hundreds of algorithms in the literature [10], no one pitch detector (so called because of the close correlation of F0 and perceived pitch) is fully accurate. Environmental noise often obscures speech periodicity, and the interaction of phase, harmonics, and spectral peaks often creates ambiguous cases where pitch detectors can make small or even large mistakes (especially in weaker sections of speech).

The basic approach to F0 estimation simply looks for peaks in the speech waveform, spaced at intervals roughly corresponding to typical pitch periods. Periods can range from 2 ms (sounds from small infants) to 20 ms (for large men), but each individual speaker usually employs about an octave range (e.g., 6-12 ms for a typical adult male). Many estimators use heuristics, such as the fact that F0 rarely changes abruptly (except when voicing starts or ceases). Since F0 is roughly independent of which phoneme is being uttered, and since the structure of formants (and related phase effects) in a voiced spectrum can obscure F0 estimation, we often eliminate from analysis the frequencies above 900 Hz (via a lowpass filter), thus retaining one strong formant containing several harmonics to supply the periodicity information. (This also simplifies processing by use of a decimated signal, which allows a lower sampling rate after lowpass filtering.) At extra cost, spectral flattening can be provided by autocorrelation methods or by LPC inverse filtering, to further reduce F0 estimation errors. Other pitch detectors do peak-picking directly on the harmonics after a Fourier transform of the speech.

5. AUTOMATIC SPEECH RECOGNITION (ASR)

Translating a speech signal into its underlying message (i.e., ASR) is a pattern recognition problem. In principle, one could store all possible speech signals, each transcribed with its corresponding text. Given today's faster computers and the decreasing cost of computer memory, one might wonder if this radically simple approach could eventually solve the ASR problem. To see that this is not true, consider the immense number of possible utterances. Typical utterances last a few seconds; at 10,000 samples/s, we potentially have, say, 2^{30,000} signals. Even with very efficient coding (e.g., 100 bps), we would still have an immense 2³⁰⁰ signals. Most of these would not be recognizable as speech, but it is impossible a priori to just consider ones that might eventually occur as an input to an ASR system (even enlisting millions of speakers talking for days would be quite insufficient).

Thus speech signals to be recognized must be processed to reduce the amount of information from perhaps 64 kbps to a much lower figure. Obvious candidates are lowrate coders, since they preserve enough information to reconstruct intelligible speech. (Higher-rate waveform coders retain too much information, which is useful for naturalness but is not needed for intelligibility; only the latter is important for ASR.)

Unlike speech coding, where the objective of speech analysis is to reproduce speech from a compact representation, ASR instead transforms speech into its corresponding textual equivalent. The direct relationship between the spectral envelope of speech and vocal tract shape (and hence to the phoneme being uttered) has led to intense use of efficient representations of spectral envelope for ASR. (The lack of a simple, direct correlation between F0 and phonemes, on the other hand, has led to F0 being largely ignored in ASR, despite its use to cue semantic and syntactic information in human speech recognition.) One difficulty has been how to extract compact yet relevant information about the envelope. Simple energy is useful, but is often subject to variations (e.g., automatic gain control, variable mouth-to-microphone distance, varying channel gain) irrelevant for phonemic distinctions; as a result, a simple energy measure is often not used for ASR, but change in energy between frames is.

LPC parameters were once popular for ASR, but they have been largely replaced by the mel-scale frequency cepstral coefficients (MFCCs) [11,12]. The term *mel-scale* refers to a frequency-axis deformation, to weight the lower frequencies more than higher ones (which is quite difficult to do in LPC analysis). This follows critical-band spacing in audition, where perceptual resolution is fairly linear below 1 kHz, but becomes logarithmic above that. The cepstrum is the inverse transform of the log amplitude of the Fourier transform of speech. The amplitude spectrum of speech in decibels is weighted via triangular filters spaced at critical bands, and then coefficients are produced via weightings from increasingly higher-frequency sinusoids (in the inverse Fourier transform step). The first 10 or so MFCCs provide a good spectral envelope representation for ASR. C0 (the first MFCC) is actually just the overall speech energy (and is often omitted from use); C1 provides a simple measure of the balance between low- and highfrequency energy (the one-period sinusoid weights low frequencies positively and high ones negatively). Higher coefficients provide the increasingly finer spectral details needed to distinguish, say, the vowels /i/ and /e/.

5.1. Timing Problems in ASR

The major difficulty for ASR is the large amount of variability in speech production. In text-to-speech synthesis, one synthetic voice may suffice, and all listeners must adjust to its accent. ASR systems, however, must accommodate different speaking styles, by storing many different speakers' patterns or by integrating knowledge about different styles. Variations occur at several levels: timing, spectral envelope, and intonation. There is much freedom in how a speaker times articulations and how exactly the vocal tract moves. In 1986, ASR commonly stored templates consisting of successive frames of spectral parameters (e.g., LPC coefficients), and compared them with those of an unknown utterance. Since utterances of the same text could easily have different numbers of frames, the alignment of frames could not simply be one-to-one. Nonlinear "dynamic time warping" (DTW) was popular because it compensated for small speaking rate variations. However, it was still computationally expensive and extended awkwardly to longer utterances; it was also difficult to improve models with additional training speech.

5.2. General Stochastic Approach

During the 1980s, the hidden Markov model (HMM) approach became the dominant method for ASR. It accepted large amounts of computation during an initial training phase in order to get a more flexible and faster model at recognition time. Furthermore, HMMs provided a mathematically elegant and computationally practical solution to some of the serious problems of variability in speech production. DTW had partially accommodated the timing problem, via its allowance of nonlinear temporal paths in the recognition search space, but was unsatisfactory for variations in pronunciation (e.g., spectral differences due to perturbations in vocal tract shape, different speakers, phonetic contexts). DTW was essentially a deterministic approach to a stochastic problem.

With HMMs, speech variability is treated in terms of probabilities. The likelihood of a speaker uttering a certain sound in a certain context is modeled via probability distributions, which are estimated from large amounts of "training data" speech. Given sufficient computer power and memory, ASR systems tend to have improved recognition accuracy as more data are implicated in the stochastic models, to make them more reliable. While computer resources are never infinite, this approach is feasible to improve systems for "speaker-independent ASR," where training speech is obtained from hundreds of different speakers, and the system then accepts input speech from all users.

For alternative "speaker-dependent" recognizers, which need training from individual users and only employ models trained on each speaker's voice at recognition time, the large amounts of training data are less feasible, given most users' reluctance to provide more than a few minutes of speech. The latter systems have better accuracy, since the HMM models are directly relevant for each user's speech, whereas speaker-independent HMMs must model much more broadly across the diversity of many speakers (such models, as a result, are less discriminative).

Most ASR is done using the ML (maximum-likelihood) approach, in which statistical models for both speech and language are estimated based on prior training data (of speech and texts, respectively). After training establishes the models, at recognition time an input speech signal is analyzed and the text with the corresponding highest likelihood is chosen as the recognition output. Thus, given a signal S, we choose text T, which maximizes the a posteriori probability P(T|S) = P(S|T)P(T)/P(S), using Bayes' rule. It is impossible to directly get good estimates for P(T|S) because of the extremely large number of possible speech signals S. Instead, we develop estimates for P(S|T) (the acoustic model) and for P(T) (the language or text model). When maximizing across possible texts T, we ignore the denominator P(S) term as irrelevant for the best choice of T. Even for large vocabularies, the number of text possibilities is much smaller than the number of speech signals; thus it is more practical to estimate P(S|T) than P(T|S). For each possible text, the acoustical statistics are obtained from speakers repeatedly uttering that text (in practice, such estimates can be obtained for small text units, such as words and phonemes, while still using speech conveniently consisting of sentences). The a priori likelihood of a text *T* being spoken is P(T), which is obtained by examining computerized textual databases.

There are efficient methods to develop such statistics and to evaluate the large number of possibilities when searching for the maximum likelihood (the search space for a vocabulary with thousands of words and for an utterance of several seconds, at typically 100 frames/s, is quite large). The "forward-backward" method examines all possible paths, summing many small joint likelihoods, while the more efficient Viterbi method looks for the single best path [11]. The latter is much faster, and is commonly used because in practice it tends to sacrifice little in recognition accuracy. When trying to discriminate among similar words, however, the ML approach sometimes fails, because it does not examine how close alternative possible texts are. In addition, HMMs treat all speech frames as equally important, which is not the case in human speech perception. Alternative methods such as maximum mutual information estimation or linear discriminant analysis are considerably more expensive and complex, but are more selective in examining the data, focusing on the differences between competing similar texts, when examining a speech input.

5.3. Details of the Hidden Markov Model (HMM) Approach

The purpose of this method is to model random dynamic behavior via a mathematical method where different "states" represent some aspects of the behavior. The basic, first-order Markov chain has several states, connected by transitions among states. The likelihood of leaving each state is modeled by a probability distribution (usually a PDF — probability density function), and each state itself is also so modeled. For speech, these two sets of PDFs attempt to model the variable timing and articulation of utterances, respectively. Each state roughly corresponds to a vocal tract shape (or more precisely to a speech spectral envelope with formants resulting from the vocal tract shape). Each transition models the likelihood that the vocal tract moves from one position to another.

The PDF for transitions from a given state usually has a simple form: a relatively high likelihood of remaining in that state (e.g., 0.8), a smaller chance of moving to the next state in the time chain (e.g., 0.15), and a yet smaller chance of skipping to the state after the next one. Since time always moves forward, we do not allow backward transitions (e.g., we go through an HMM modeling a word, starting with its first sound and proceeding to its last sound). Since each state models roughly a vocal tract shape (or more often an average of shapes), we stay in each state for several 10-ms frames (hence the high self-loop likelihood). Allowing an occasional state to be skipped accounts for some variability in speech production, especially for rapid, unstressed speech. For example, the training speech may be clear and slow, thus creating states that need not always be visited in later (perhaps fast) test speech.

The transition PDF thus described leads to an exponential PDF for the duration of state visits, which is an inaccurate model for actual phoneme durations. There is too much bias toward short sounds. For instance, very few phonemes last only 1-2 frames. Some more complicated HMM approaches allow direct durational modeling, but at the cost of increased complexity.

In practice, every incoming speech signal is divided into successive 10-ms frames of data for analysis (as in speech coding). During the training phase, many frames are assigned to each given HMM state, and the state's PDF is simply the average across all such frames. Similarly, the PDF describing which state B follows any given state A in modeling the dynamics of the speech simply follows the likelihood of moving to a nearby vocal tract shape (B) in one frame, given that the previous speech frame was assigned to state A. This simple approach in which the model takes no direct account of the history of the speech (beyond one state in the past) is called a first-order Markov model. It is clear, when dealing with typical 10ms frames, that there is certainly significant correlation across many successive frames (due to coarticulation in vocal tract movements). Thus the first-order assumption is an approximation, to simplify the model and minimize computer memory. Attempts to use higher-order models have not been successful because of the immense increase in complexity needed to accommodate reasonable amounts of coarticulation. The HMM is called "hidden" because the vocal tract behavior being modeled is not directly observable from the speech signal (if the input to an ASR system were X rays of the actual moving vocal tract during speech, we could use direct Markov models, but this is totally impractical).

Gaussian PDFs are normally used to specify each state in an HMM. Such a PDF for an *N*-dimensional feature vector \mathbf{x} for a spoken word assigned index *i* is

$$P_{i}(\mathbf{x}) = (2\pi)^{-N/2} |\mathbf{W}_{i}|^{-1/2} \exp\left[\frac{-(\mathbf{x} - \mu_{i})^{\mathrm{T}} \mathbf{W}_{i}^{-1}(\mathbf{x} - \mu_{i})}{2}\right]$$
(1)

where \mathbf{W}_i is the covariance matrix (noting the individual correlations between parameters), $|\mathbf{W}_i|$ is the determinant of \mathbf{W}_i , and μ_i is the mean vector for word *i*. Most systems use a fixed W matrix (instead of individual \mathbf{W}_i for each word) because (1) it is difficult to obtain accurate estimates for \mathbf{W}_i from limited training data, (2) using one W matrix saves memory and computation, and (3) \mathbf{W}_i matrices are often similar for different words. If one chooses parameters that are independent (i.e., no relationship among the numbers in the vector \mathbf{x}), the W matrix simplifies to a diagonal matrix, which greatly simplifies calculation of the PDF (which must be repeatedly done for each speech frame and for each HMM). Thus many ASR systems assume (often without adequate justification, other than reducing cost) such a diagonal W. Unless an orthogonalization procedure is performed (e.g., a Karhunen-Loeve transformation-itself quite costly), most commonly used feature sets (LPC coefficients, MFCCs) have significant correlation.

The elements along the main diagonal of **W** indicate the individual variances of the speech analysis parameters. The use of \mathbf{W}_i^{-1} in the Gaussian PDF notes that those features with the smallest variances are the most useful for discriminating sounds in ASR. Parameter sets that lead to small variances and widely spaced means μ_i are best, so that similar sounds can be consistently discriminated.

The Gaussian form for the state PDF is often appropriate for modeling many physical phenomena; the idea follows from basic stochastic theory: the sum of a large number of independent, identically distributed random variables resembles a Gaussian. Natural speech, coming from human vocal tracts, can be so treated, but only for individual sounds. If we try to model too many different vocal tract shapes with one HMM state (as occurs in multispeaker, or context-independent ASR), the Gaussian assumption is much less reasonable (see text below for the discussion of mixtures of Gaussian PDFs).

We create HMMs to model different units of speech. The most popular approaches are to model either phonemes or words. Phonemic HMMs require fewer states than word-based HMMs, simply because phonemes are shorter and less phonetically varied than words. If we ignore coarticulation with adjacent sounds, many phonemes could be modeled with just one state each. Inherently dynamic phonemes, such as stops and diphthongs, would require more states (e.g., a stop such as /t/ would at least need a state for the silence portion, a state for the explosive release, and probably another state to model ensuing aspiration). There is a very practical issue of how many states are appropriate for each HMM. There is no simple answer; too few states lead to diffuse PDFs and poor discriminability (especially for similar sounds such as /t/ and /k/), due to averaging over diverse spectral patterns. Too many states lead to excessive memory and computation, as well as to "undertraining," where there are not enough speech data in the available training set to provide reliable model parameters.

Consider the (very practical) task of simply distinguishing "yes" versus "no" (or perhaps just the 10 digits: 0, 1, $2, \ldots 9$); it is efficient to create an HMM for every word in the allowed vocabulary. There will be several states in each HMM to model the sequence of phonemes in each word. HMMs work best when there is a rough correspondence between the number of states and the number of distinguishable sounds in the speech unit being modeled. If the model creation during the training phase is done well, this allows each state to have a PDF with minimal variance. At recognition time, such tight PDFs will more easily discriminate words with different phoneme sequences.

As the size of the allowed vocabulary increases, however, it becomes much less practical to have one or more models for every word. Even in speaker-independent ASR, where we could ask thousands of different people to furnish speech data to model the many tens of thousands of words in any given language such as English, the memory and search time needed for word-based HMMs goes beyond current computer resources. Thus, for vocabularies larger than 1000 words, most systems employ HMMs that model phonemes (or sometimes "diphones," which are truncated sequences of phoneme pairs, used to model coarticulation during transitions between two phonemes). Diphones are obtained by dividing a speech waveform into phoneme-sized units, with the cuts in the middle of each phone (thus preserving in each diphone the transition between adjacent phonemes). In text-to-speech applications, concatenating diphones in a proper sequence (so that spectra on either side of a boundary match) usually yields smooth speech because the adjacent sounds at the boundaries are spectrally similar; e.g., to synthesize straight, the six-diphone sequence /#s-st-tr-re-et-t#/ would be used (# denoting silence).

Using phonemic HMMs reduces memory and search time, as well as allowing a fixed set of models, which does not have to be updated every time a word is added to the vocabulary. If the models are "contextindependent," however, their discriminability is small. When averaging over a wide range of phonetic contexts, the states modeling the initial and final parts of each phoneme have varying spectral patterns and thus broad PDFs of poorer discriminability.

Nowadays, more advanced ASR systems employ "context-dependent" phonemic HMMs (e.g., diphone models), where each phoneme is assigned many HMMs, depending on its immediate neighbors. A simple and common (but expensive) technique uses triphone models, where for a language with N phonemes we have N^3 models; each phoneme has N^2 models, one for each context (looking before and after by one phoneme). There is much evidence from the speech production literature about the effects of coarticulation, which are significant even beyond a triphone window. (A diphone approach is a compromise between a simple context-independent scheme and a complex triphone method; it has N^2 models.) For English, with about 32 phonemes, the triphone method leads to tens of thousands of HMMs (which need to be adequately analyzed and stored in the training phase, and all must be examined at recognition time). Many systems adopt a compromise by grouping or clustering similar models according to some set of phonetic rules (e.g., the decision-tree approach). If the clustering is done well, one could even expand the analysis window beyond triphones, to accommodate further coarticulation (e.g., in the word "strew," lip rounding for the vowel /u/ affects the spectrum of the initial /s/; thus a triphone model for /u/looking leftward only at the neighbor /r/ is inadequate).

5.4. Language Models for ASR

Going back only a decade or so, we find much less successful ASR methods that concentrated only on the acoustic input to make decisions. It was thought that all the required information to translate speech to text could be found in the speech signal itself; thus no cognitive modeling of the listener seemed necessary. Some surprisingly simple experiments that modeled some aspects of language changed that naive approach, with positive results in recognition accuracy. Indeed, it can be easily observed that words in speech (as in text) rarely occur in random order. Both syntactically and semantically, there is much redundancy in word order. For example, when talking about a dog, one may well find the semantically related words "small" or "black" immediately prior to "dog." As for syntax, there are many restrictions on word sequences, such as the common structure of article+adjective+noun for English noun phrases, and strict order in verb phrases such as "may not have been eaten."

Thus researchers developed models for language, in which the likelihood of a given word occurring after a prior word sequence is evaluated and used in the decision process of speech recognition. The most common approach is that of a trigram language model, where we estimate the probability that any given word in text (or speech) will follow its preceding two words (written or spoken). Textual redundancy in English (and indeed many languages) goes well beyond a trigram window of analysis, but practical issues of computer memory and the availability of training data have so far limited most models to a three-word range. Training for language models has been almost exclusively done using written texts (and rarely transcriptions of speech), despite the inappropriateness for ASR of written compositions (i.e., most speech is spontaneous, rarely from written texts), owing to the cost (and often limited accuracy) of transcribing large amounts of speech. Researchers find it much easier to use the many databases of text available today, despite their shortcomings.

The major advantage of models such as trigrams is that they incorporate diverse aspects of practical text redundancies automatically (no complex semantic or syntactic analysis is needed). This reflects the current premature state of natural language processing (e.g., the continuing difficulties in automatic language translation). This lack of intelligent analysis in trigram models, however, leads to serious inefficiencies. As the allowed vocabulary increases in size, for applications not limited to specific topics of conversation, the availability of sufficient text to obtain reliable statistics is a major problem. Depending on what is counted as a word (e.g., do "eat, eats, eating, eaten" count as four individual words?), there are easily hundreds of thousands of words in English; taking a very conservative estimate of 10⁵ words, there are 10^{15} trigrams, which strains current computer capacities, or at least significantly increases costs. As a result, many acceptable trigrams (and even many bigrams and unigrams) are not observed in any given training text (even large ones containing millions of words). The usual estimation of probabilities simply divides the number of occurrences of each specific trigram (or bigram) by the total number of occurrences of each word being modeled in the training text. For sequences occurring frequently in training texts, the estimates are reasonable, at least for the purposes of modeling (and recognizing speech from) additional text from the same source. One may readily create different language models for different types of text, such as those for physicians, lawyers, and other professionals, corresponding to specific applications.

A serious problem with the approach above is how to handle unobserved sequences (e.g., the very large number of three-word sequences not found in a given training text, but nonetheless permitted in the language). Assigning them all a probability of zero is quite inappropriate, since such word sequences would then never appear in a speech recognizer's output, even if they were actually spoken. One solution simply assigns all unseen sequences the same tiny probability, and reduces the probabilities of all observed sequences by a compensatory amount (to ensure that the total probability remains equal to one). This is somewhat better than arbitrarily assigning zero likelihoods, but treats all unseen sequences as equally likely, which is a very rough and poor estimate. A more popular way is the "backoff" approach, in which the overall assigned likelihood for a word is a weighted combination of trigram, bigram, and unigram probabilities. The weightings are appropriately adjusted for unseen sequences. If no trigram (or bigram) estimate is available, one simply assigns its corresponding weighting a value of zero, and uses existing

(bigram and) unigram statistics. Of course, words that never appear at all in a training set have no statistics to which to back off; their likelihoods must be estimated another way.

While the trigram method has considerable power as a language model for ASR, there is also considerable waste. The memory for trigram statistics for large-vocabulary applications is very large, and many of the probabilities are poorly estimated due to insufficient training data. It is more efficient if one clusters similar words together, thus reducing memory and making the fewer statistics more reliable. One extreme case is the tri-PoS approach, where all words are classified by their syntactic category (e.g., part of speech, such as noun, verb, or preposition). Depending on how many categories are used, the statistics can be very reduced in size, while still retaining some powerful syntactic information to predict common wordclass sequences. Unfortunately, the semantic links among adjacent words are largely lost with this purely syntactic language model. Owing to the reduced model size, one can easily extend the tri-PoS method to windows of more than three words, or one can extend the word categories to include semantic labels as well. Word-class language models using perhaps hundreds of hybrid syntacticsemantic classes may eventually be a good compromise approach between the too simple tri-PoS method and the huge trigram approach.

5.5. Practical Issues in ASR — Segmentation

In addition to the serious ASR issues of computer resources and of speech variability, there is another aspect of recognizing speech that makes the task more difficult than synthesizing speech: segmentation. In the case of automatic speech synthesis, the computer's task is easier for two reasons: (1) the input text is clearly divided into separate words, which simplifies processing; and (2) the burden is placed on the human listener to adapt to the weaknesses of the computer's synthetic speech. In ASR, on the other hand, the machine must adapt to each user's different style of speech, and there are few clear indications to boundaries in a speech signal. In particular, acoustic cues to word boundaries in speech are very few. Automatic segmentation of a long utterance into smaller units (ideally into words) is very difficult, but is effectively necessary to reduce computation and lower error rates. One is tempted to exploit periods of silence for segmentation, but they are unreliable cues to linguistic boundaries. Long pauses are usually associated with sentence boundaries, but they often occur within a sentence and even within words (e.g., in hesitations). Short pauses are easily confused with phonemic silences (i.e., the vocal tract closures during unvoiced stops).

Segmenting speech into syllabic units is easier, because of the typical rise and fall of speech energy between vowels and consonants. However, many languages (e.g., English) allow many different types of syllables (ranging from ones with only a vowel to ones with several preceding and ensuing consonants). Many languages (e.g., Japanese) are much easier to segment because of their consistent alternation of consonants and vowels. Segmenting speech into phonemes is also difficult if the language allows sequences of similar phonemes, such as vowels, or large variations in phonemic durations (e.g., English allows severe reduction of unstressed phonemes, such that recognizers often completely miss brief sounds; at the other extreme, long vowels or diphthongs can often be misinterpreted as containing multiple phonemes).

To facilitate segmentation, many commercial recognizers still require speakers to adopt an artificial style of talking, pausing briefly (at least a significant fraction of a second) after each word. Silences of more than, say, 100 ms should not normally be confused with (briefer) stop closures, and such sufficiently long pauses allow simple and reliable segmentation of speech into words.

In order of increasing recognition difficulty, four styles of speech are often distinguished: *isolated-word* or *discrete-utterance* speech, *connected-word* speech, *continuous*-read speech, and normal conversational speech. The last two categories concern continuous-speech recognition (CSR), which requires little or no adaptation of speaking style by system users. CSR allows the most rapid input (e.g., 150-250 words/min), but is the most difficult to recognize. For isolated-word recognition, requiring the speaker to pause after each word is very unnatural for speakers and significantly slows the rate at which speech can be processed (e.g., to about 20-100 words/min), but it clearly alleviates the problem of isolating words in the input speech signal.

Using word units, the search space (in both memory and computation) is much smaller than for longer utterances, and this leads to faster and more accurate recognition. However, few speakers like to talk in "isolated word" style. There remains also the serious issue of "endpoint detection," where the beginning and end times for speech must be determined; against a background of noise, it is often difficult to discriminate weak sounds from the noise. Endpoint detection is more difficult in noise: speaker-produced (lip smacks, heavy breaths, mouth clicks), environmental (stationary: fans, machines, traffic, wind, rain; nonstationary: music, shuffling paper, door slams), and transmission (channel noise, crosstalk). The large variability of durations and amplitudes for different sounds makes reliable speech versus silence detection difficult; strong vowels are easy to locate, but the boundaries between weak obstruents and background noise are often poorly estimated. Typically, endpoint location uses energy as the primary measure to cue the presence of speech, but also employs some spectral parameters. Clear endpoint decisions are required very often in isolated-word speech. For these various reasons, most current research focuses on more normal "continuous" speech.

Stochastic modeling has an important place in speech processing, given the large amount of variability in speech production. Humans are indeed incapable of producing exactly the same utterance twice. With effort, they can likely make utterances sound virtually the same to listeners, but at the sampling and precision levels of automatic speech analyzers, there are always differences, which result in parameter variations. If one selects ASR parameters well and such variation is minimal, system performance can be high with proper stochastic models. However, all too often, variability is large and goes beyond the modeling capability of current ASR systems. Future ASR must find a way to combine deterministic knowledge modeling (e.g., the "expert system" approach to artificial intelligence tasks) with well-controlled stochastic models to accommodate the inevitable variability. Currently, the research pendulum has swung far away from the expertsystem approach common in the mid-1970s.

5.6. Practical Issues in ASR: Noise

In many applications, the speech signal is subject to various distortions before it is received by a recognizer [13]. Noise may occur at the microphone (e.g., in a telephone booth on the street) or in transmission (e.g., fading over a portable telephone). Noise leads to decreased accuracy in the parameters during speech analysis, and hence to poorer recognition, especially (as is often the case) if there is a mismatch in conditions between the training and testing speech. It is very difficult (and/or expensive) to anticipate all distortion possibilities during training, and thus ASR accuracy certainly decreases in such cases.

As an attempt to normalize across varying conditions, ASR systems often calculate an average parameter vector over several seconds (e.g., over the entire utterance), and then subtract this from each frame's parameters, applying the net results to the recognizer. This approach takes account of slow changes in average energy as well as the filtering effects of the transmission microphone and channel, which usually change slowly over time. Such "mean subtraction" is simple and useful in noisy conditions, but can delay recognition results if we must determine the mean for several seconds of speech and only then start to recognize the speech. A similar method called RASTA (RelAtive SpecTrAl) uses a highpass filter (with a very low-frequency cutoff) to eliminate very slowly varying aspects of a noisy speech signal, as being specific to the transmission channel and irrelevant to the speech message.

More traditional speech enhancement techniques can also be used in ASR [14]. These include spectral subtraction (or related Wiener filtering), where the average amplitude spectrum observed during signal periods that are estimated to be devoid of actual speech is subtracted from the spectrum of each speech frame. In this way, stationary noise can be somewhat suppressed. Frequency ranges dominated by noise are thus largely eliminated from consideration in the analysis to determine relevant parameters. The output of speech enhancers sounds more pleasant, but intelligibility is not enhanced by this removal of noise.

Sometimes used for speech enhancement, another approach called "comb filtering" estimates F0 in each frame and suppresses portions of the speech spectrum between the estimated harmonics. Again, the enhanced output speech sounds more pleasant, but this method requires a reliable F0 detector, and the resulting comb filter must adapt dynamically to the many changes in pitch found in normal speech.

Much more powerful speech enhancement is possible if several microphones are allowed to record in a noisy environment. For example, in a noisy plane cockpit, one microphone close to the pilot's lips would capture a signal containing speech plus some noise, while another microphone outside the helmet would capture a version of the noise corrupting the initial signal. Using adaptive filtering techniques very similar to those for echo cancellation at 2/4-wire junctions in the switched telephone network, one can improve very significantly both the quality and the intelligibility of such speech.

5.7. Artificial Neural Networks (ANNs)

Since 1990, in another attempt to solve many problems of pattern recognition, significant research has been done in the field of artificial neural networks. These ANNs simulate very roughly the behavior of neurons in the human central nervous system. Given a signal (e.g., speech) to recognize, each node of an ANN accepts a set of signal-based inputs, computes a linear combination of the these values, compares the result against a threshold, and provides a binary output (1 if the linear combination exceeds the threshold, 0 otherwise). By combining such nodes in a network, quite complicated decision spaces can be constructed, which have proven useful in numerous pattern recognition applications, including speech recognition [15]. For ASR, the ANN typically accepts a long vector of L parameters (e.g., for 10 MFCCs per frame and a 100-frame utterance, one has a vector of L = 1000 dimensions), as emitted by the usual speech analysis methods discussed above, and these Lnumbers provide the input to a set of M nodes (each individual input value may be input to several nodes). Each node thus makes a linear combination of a selected set of values from the input vector. The conceptually simplest ANN would assign one output node to each possible output of the recognition process (e.g., if the utterance must be one of the 10 spoken digits, M = 10, and one tries to specify the weights in the ANN such that, on any given trial, only one of the *M* outputs will be 1; the label on that output node provides the textual output. Thus, with proper training of the ANN to choose the appropriate weights, when one says "six" and the L resulting analysis parameters are fed through the ANN, ideally only one of the 10 output nodes (i.e., that corresponding to "six") will show a 1 (the others showing 0). This assumes, however, that each of the allowed vocabulary words corresponds to a simple distribution in L-dimensional acoustic space, such that basic hyperplanes can partition the space into 10 separable clusters (and furthermore that a training algorithm looking at many utterances of the 10 digits can determine the correct weights). It is very rare that these assumptions are exactly true, however. To adapt to practical speech cases, a partial solution has been to add extra layers of nodes to the ANN. We allow the number M_2 of nodes in the second layer to be larger than M, and allow the binary outputs of those M_2 nodes to pass through another layer of weights and N nodes. Such a double-layer ANN can partition the speech space into the much more complex shapes that correspond to practical systems. For speech, in general, it appears that a threetier system is needed; the original input speech-parameter vector provides the numbers for the lowest layer, the information propagates through three levels of weightings (two "hidden" layers of nodes) to finally appear at the output layer (with one node for each textual label).

For static patterns (e.g., simple steady vowels), ANNs have provided excellent classification when properly trained. However, general ASR must handle dynamic speech, and thus the practical success of ANNs has been more limited. So-called time-delay ANNs allow complicated feedback within ANN layers to try to account for the fact that small delays (variability in timing) in speech signals occur often in human speech communication and do not hinder perception. In a basic ANN, however, shifting the set of inputs can cause large changes in the ANN outputs. As a result, for ASR, ANNs have typically found application, not as a replacement for HMMs, but as an additional aid to the basic HMM scheme (e.g., ANNs can be used in the training phase for better estimation of HMM parameters).

6. SPEAKER VERIFICATION

We have shown speech analysis to be useful in coding, where speech is regenerated after efficient compression, and in ASR, where speech is converted to text. Another application for speech analysis is speaker verification, where one determines whether speakers are who they claim to be. Fingerprints or retinal scans are often more accurate in identifying people than speech-based methods, but speech is much less intrusive and is feasible over the telephone. Many financial institutions, as well as companies furnishing access to computer databases, would like to provide automatic customer service by telephone. Since personal number codes (keyed on a telephone pad) can easily be lost, stolen, or forgotten, speaker verification may provide a viable alternative. The required decision process can be much simpler than for ASR, since a given speech signal is converted into a simple binary digit ("yes," to accept the claim, or "no," to refuse it). However, the accuracy of such verification has only relatively recently reached the point of commercialization.

In ASR, for speech signals corresponding to the same spoken text, any variation due to different speakers is viewed as "noise" to be either (1) (ideally) eliminated by speaker normalization or (2) (more commonly) accommodated through models based on a large number of speakers. When the task is to verify the person talking, rather than what that person is saying, the speech signal must be processed to extract measures of speaker variability instead of segmental acoustic features. What to look for in a speech signal for the purpose of distinguishing speakers is much more complex than for ASR. In ASR we look for spectral features to distinguish different fundamental vocal tract positions, and we can exploit language models to raise accuracy.

One common speaker verification approach indeed employs ASR techniques, with the exception of labeling models with speaker names instead of word or phoneme labels. When two speakers utter the same word or phoneme, spectral patterns are compared to see which speaker provides the more precise match. In ASR, competing phoneme or word candidates can often be distant in spectral space, making the choice relatively easy. In speaker verification, however, many people have quite similar vocal tract shapes and speaking styles; using analysis over many frames, however, renders the task feasible.

For speech recognition, much is known about the human speech production process, which links a text and its phonemes to the spectra and prosodics of a corresponding speech signal. Each phoneme has specific articulatory targets, and the corresponding acoustic events have been well studied (but still are not fully understood). For speaker verification, on the other hand, the acoustic aspects of what characterizes the principal differences between voices are difficult to separate from signal aspects that reflect ASR. There are three major sources of variation among speakers: differences in vocal cords and vocal tract shape, differences in speaking style (including variations in both target positions for phonemes and dynamic aspects of coarticulation such as speaking rate), and differences in what speakers choose to say. Automatic speaker verification exploits only the first two of the variation sources, examining low-level acoustic features of speech, since a speaker's tendency to use certain words and syntactic structures (the third source) is difficult to exploit (e.g., impostors could easily mimic a speaker's choice of words, but would find simulating a specific vocal tract shape or speaking style much more difficult).

Unlike the relatively clear correlation between phonemes and spectral resonances, there are no evident acoustic cues specifically or exclusively dealing with speaker identity. Most of the parameters and features used in speech analysis contain information useful for the identification of both the speaker and the spoken message. The two types of information, however, are coded quite differently in the speech signal. Unlike ASR, where decisions are made for every phoneme or word, speaker verification requires only one decision, based on parts or all of a test utterance, and there is no clear set of acoustic cues that reliably distinguishes speakers. Speaker verification typically utilizes long-term statistics averaged over whole utterances or exploits analysis of specific sounds. The latter approach is common in "text-dependent" applications where utterances of the same text are used for both training and testing.

There are two classes of errors in speaker verification: false rejections and false acceptances. A false rejection occurs when the system incorrectly rejects a true speaker. A false acceptance occurs when the system incorrectly accepts an impostor. The decision to accept or reject usually depends on a threshold: if the distance between a test speech pattern and a reference pattern exceeds the threshold (or equivalently, using a PDF for the claimed speaker, the evaluated probability value is too low), the system rejects a match. Depending on the costs of each type of error, verification systems can be designed to minimize an overall penalty by biasing the decisions in favor of less costly errors. Low thresholds are generally preferred because false acceptances are usually more expensive (e.g., admitting an impostor to a secure facility might be disastrous, while excluding some authorized customers is often merely annoying). System parameters are often adjusted so that the two types of error occur equally often (*equal error rate condition*).

One popular method is the Gaussian mixture model [16], which follows the popular modeling approach for ASR, but eliminates separate HMMs for each phoneme as unnecessary. It is essentially a one-state HMM, but allows the PDF to be complicated. In ASR, we can often justify that each state's PDF may be approximated as approaching a Gaussian distribution (especially for triphone HMMs, where context is well controlled). When merging all speech from a speaker into one PDF, however, that PDF is far from Gaussian. To retain the simple statistics of Gaussians (complete specification with only the mean and variance), however, both ASR and speaker verification often allow a state's PDF to be a weighted combination of many Gaussian PDFs. Using the same set of Gaussians for all speakers, each speaker is characterized simply by the corresponding set of weights. Evaluation is thus quite rapid, even when all the frames of an utterance contribute to the overall speaker decision.

Such an approach ignores dynamic speech behavior (as well as ignoring intonation, as do most ASR systems), and further assumes that both training and testing utterances will be roughly similar in phonetic composition. This latter assumption is ensured when doing "textdependent" verification, where the speaker is asked to utter words already made known to the system during the training phase; for instance, a commercial system may train on two-digit numbers (e.g., "74") and then ask each candidate speaker to utter a few such numbers at test time. False rejections are much lower for these systems than for "text-independent" verifiers, although the latter are sometimes needed for forensic work (e.g., there is no text control in wiretapped conversations). The former, however, run the risk of an impostor playing a recording of the desired speaker (over the telephone, or if there is no camera surveillance at a secure facility). Such impostors, of course, elevate false-acceptance rates with "textindependent" verifiers (which have lower accuracy) as well, but impostors would have difficulty anticipating the requested text that customers are asked to speak, in the case of the much larger vocabulary in the latter systems.

In any pattern recognition task (ASR included), training and test data must be kept separate (i.e., testing on data already used for training artificially raises performance to levels that will usually not be replicated in practical testing situations), but this is even more important in the case of speaker verification. Speakers tend to vary their style of speech over time (e.g., morning versus evening, Monday-Saturday, healthy-ill). Training a speaker verifier in only one session, and then testing a few months later will usually not give good results. Ideally, speaker verification needs months of training (or at least an adaptive system, which allows for periodic updates while being used). If the same utterances are used to train and test a system, a high degree of accuracy is due to the model parameters often being heavily tuned toward the training data. For good results, the training data must be sufficiently diverse to represent many different possible future input speech signals. With shared training and test data, it is difficult to know whether the system

has been designed to take advantage of specific speech or speaker characteristics that may not be repeated in new data. Given K utterances per speaker as data, one common procedure trains its system using K - 1 as data and one as test, but repeats the process K times treating each utterance as test once. Technically, this "leave one out" method designs K different systems, but this method verifies whether the system design is good using a limited amount of data, while avoiding the problems of common train and test data.

7. TEXT-TO-SPEECH SYNTHESIS (TTS)

Our final speech processing application concerns the automatic generation of speech from text [17-19]. Here, speech processing occurs both in the development stage of the system, when spoken units from one or more speakers are recorded, and again at synthesis time, when selected stored units are concatenated to produce the synthetic output voice. The units involved can range from full sentences (e.g., as in talking cars and ovens) to shorter phrases and words (e.g., telephone directory assistance) to phonemes (e.g., for unlimited text applications).

We distinguish true TTS systems, which accept any input text in a chosen language (this could include new words and typographical errors), from "voice response systems," which accept a very limited vocabulary and are essentially voice coders of much simpler complexity (but suffer from inflexibility). Commercial synthesizers have now become much more widespread owing to both advances in computer technology and improvements in the methodology of speech synthesis.

The simplest applications requiring only small vocabularies are speech coders that merely play back the speech when needed. This is impossible for general TTS applications, because no one training speaker can provide all possible utterances. For TTS, small units (typically, phonemes or diphones) are concatenated to form the synthetic speech, and significant adjustments must be done at unit boundaries to avoid unacceptable disjointed (jumpy) speech. The quantity and complexity of such adjustments vary in indirect proportion to the unit size, which in turn also varies inversely with quality (e.g., general text-tospeech using smaller units sounds less natural).

The critical issues for synthesizers concern tradeoffs among the conflicting demands of simultaneously maximizing speech quality, while minimizing memory space, algorithmic complexity, and computation speed. While simple TTS is possible in real time with low-cost hardware, there is a trend toward using more complex programs (tens of thousands of lines of code, and megabytes of storage). Real-time TTS systems produce speech that is generally intelligible, but lacks naturalness.

Most synthesizers reproduce speech of bandwidth ranging from 300-3000 Hz (e.g., for telephone applications) to 100-5000 Hz (for higher quality). A spectral range up to 3 kHz is sufficient for vowel perception because vowels are adequately specified by the lowest three formants. The perception of some consonants, however, is slightly impaired when energy in the 3-5-kHz range is omitted. Frequencies above 5 kHz help improve speech naturalness but seldom aid speech intelligibility. If we assume that a synthesizer reproduces speech up to 4 kHz, a rate of 8000 samples/s is needed. Since linear PCM requires 12 bits/sample for toll-quality speech, storage rates near 100 kbps result, which are acceptable only for synthesizers with very small vocabularies.

The memory requirement for a simple synthesizer is often proportional to its vocabulary size. Continued decreases in memory costs have allowed the use of very simple TTS systems. Nonetheless, storing all possible speech waveforms for synthesis purposes (even with efficient coding) is impractical for TTS. The sacrifices usually made for large-vocabulary synthesizers involve simplistic modeling of spectral dynamics, vocal tract excitation, and intonation. Such modeling usually causes quality limitations that are the primary problems for current TTS research.

7.1. The Steps in Producing Speech from Text

TTS requires several steps to convert the linguistic textual message into an acoustic signal. The linguistic processing is, in a sense, an inverse of the procedures for ASR. First, a "preprocessing" stage "normalizes" the input text so that it is a series of spelled-out words (retaining any punctuation marks as well). All abbreviations and digits are converted to words, typically via a lookup table or simple programs (sophisticated systems however might distinguish ways of saying "\$19.99" and "1999"); word context may be necessary to handle cases such as "St. Peter St." The words are then converted into a string of phonemes (and basic intonation parameters), usually via a combination of a dictionary and pronunciation rules. Reduced memory costs have led to use of large dictionaries containing all the words in a given language and their phonemic pronunciations. With a dictionary, additional relevant information can be readily available, including lexical stress (which syllable in each word is stressed), part of speech, and even semantics.

The alternative approach of letter-to-phoneme rules is useful to handle new, foreign, and mistyped words (cases where simple dictionaries fail). Complex languages such as English (derived from both Romance and Germanic languages) require many hundreds of these rules to pronounce letter sequences correctly (e.g., consider "-ough-:" rough, cough, though, through, thought, drought). Languages are often much simpler phonetically; for instance, Spanish employs just one rule per letter. In all cases, developing TTS capability for a language requires establishing a dictionary and/or pronunciation rules. Hence commercial TTS exists only for about 20 of the world's languages, whereas many commercial ASR systems (those based on word units and without a language model) can handle virtually all languages, since most languages employ similar versions of stops, fricatives, vowels, and nasals.

The next TTS step is intonation specification, determining a duration and amplitude for each phoneme, as well as a complete F0 pattern for the utterance. This is much more difficult than the prior two TTS steps, and requires a syntactic and semantic analysis of the input text. Most languages use intonation in complex ways to cue many aspects of speech communication. Many languages (including English) stress only a small number of syllables in any utterance. This involves not just simple lexical stress from a dictionary but also a judgment about the semantic importance of the words in a sentential context; for synthetic speech comparable to that of humans, this would require a natural-language processor typically beyond current capabilities. People often cue syntactic structure via intonation; for example, in English, long word phrases often start with an F0 rise and end with a fall. Questions in many languages are cued with a large final F0 rise if they request a yes/no answer (but not questions with the "wh-words": what, when, why, who, how). Tone languages (e.g., Chinese, Thai) employ four or five different F0 patterns on syllables to distinguish different words with the same phoneme sequence. Finally, speech uttered with emotion often changes intonation significantly.

In the last TTS step, speech units are concatenated using the specified intonation, and adjustments are made to the model parameters at unit boundaries. Few such manipulations are needed for phrasal concatenation, but smaller units require at least smoothing of all parameters across the boundary for several frames. This is relatively straightforward when the units contain spectral parameters (e.g., LPC coefficients or formants), although improved quality occurs with more complicated smoothing rules. Storing small units with waveform coding (e.g., log PCM or ADPCM) is often not suitable (despite the higher general quality of such speech) here because smoothing the available parameters does not approximate well the actual coarticulation and F0 manipulation found in human speech production.

Smoothing of parameters at the boundaries between concatenated units is most important for short units (e.g., phones) and decreases in importance for larger units owing to fewer boundaries. Smoothing is much simpler when the joined units approximately correspond at the boundaries. Since diphone boundaries link spectra from similar sections of two realizations of the same phoneme, smoothing rules for their concatenation are simple. Systems that link phones, however, must use complex smoothing rules to represent coarticulation in the vocal tract. Not enough is understood about coarticulation to establish a complete set of rules to describe how the spectral parameters for each phone are modified by its neighbors. Diphone synthesizers circumvent this problem by storing parameter transitions from one phone to the next, since coarticulation influences primarily the immediately adjacent phones. However, since coarticulation often extends over several phones, using only average diphones or those from a neutral context leads to lower-quality synthetic speech. Improved quality is possible by using multiple diphones dependent on context, effectively storing "triphones" of longer duration (which may substantially increase memory requirements). Some coarticulation effects can be approximated by simple rules, such as lowering all resonant frequencies during lip rounding, but others such as the undershoot of phoneme target positions (which occurs in virtually all speech) are much harder to model accurately.

It is in this last TTS step that the system yields poorer quality than many speech coders, because TTS is often forced to employ synthetic-quality coding techniques. The traditional excitation for vocal tract filters in TTS (either LPC or formant-based approaches, which constitute most commercial methods) is a simple train of periodic pulses (for voiced speech) or white noise (a randomnumber generator) for unvoiced speech. The combination of oversimplified excitation and the limited modeling accuracy of the LPC or simple formant models leads to intelligible, but slightly unnatural, synthetic speech.

Since the late 1980s, some commercial systems have been successful in concatenating waveform-coded small units, with limited amounts of perceptually annoying spectral jumps at unit boundaries; for instance, the PSOLA (pitch-synchronous overlap and add) method outputs successive smoothed pitch periods [20]. While such speech can sound more natural, it is inflexible in producing alternative voices. It is typically based on one speaker uttering a large inventory of diphones; for a language such as English with about 32 phonemes, about 1000 diphones must be uttered and in a uniform fashion. One cannot simply adjust some synthesizer parameters here to get other synthetic voices (as is possible with formant synthesizers).

8. CURRENT TECHNOLOGY

Today's speech coders deliver toll quality (i.e., equivalent to the analog telephone network) at 8 kbps with minimal amounts of delay (and even at 4 kbps, if delay is less of an issue). The favored approach is CELP, for which there are several standards accepted internationally (e.g., in digital cellular telephony). The digital links in most telephone networks still employ simpler 64 kbps log PCM coding; 24–32 kbps ADPCM and delta modulation are also still popular because of their relative simplicity. We are still far from an ultimate limit of perhaps 100 bps to code speech. Despite the reducing cost of computer memory and speed, further research in coding is needed because of the increasing use of wireless telephony, where limited bandwidth is very much an issue.

The lack of formal standards for human-machine applications of speech (i.e., synthesis and recognition) hinders comparison of commercial systems. Several companies offer unlimited text-to-speech for several languages (typically the major European languages, plus Japanese and Chinese). Such synthetic speech is largely intelligible, but is easily discerned as synthetic and lacking naturalness. All synthesizers suffer from our lack of understanding of the complex relationships between text and intonation. Synthesizers are clearly increasing in use, but wider public acceptance awaits further improvements in quality.

Speech recognizers are also increasing in use, but their severe limitations (compared to human speech perception) have also hindered wider acceptance. The need to pause between words, restrict the choice of words, and/or do prior training, as well as frequent recognition errors, have significantly limited the use of ASR. Systems eliminating all these restrictions (i.e., continuous, speakerindependent, large-vocabulary ASR) still suffer from high cost and frequent errors, especially if they are used in noisy or telephone environments, and such conditions occur often in practical applications. Progress in ASR has been attributed more to general improvements in computers and to the wider availability of training data, than to algorithmic breakthroughs. The basic HMM approach using MFCCs was developed largely before 1985. Even more recent additions, such as delta coefficients, mean subtraction, Gaussian mixtures, and language models, have been in wide use since 1990.

Future systems will likely integrate more structure into the stochastic approach. It is clear that the expert-system approach to ASR common in the early 1970s will never replace stochastic methods, for the simple reason that individual human phoneticians can never assimilate enough information from hundreds of hours of speech, in ways that probabilistic computer models can improve with larger amounts of training data. The extremely simple stochastic models in current widespread ASR use, however, are too unstructured and allow too much freedom (similar complaints hold for recent neural network approaches to ASR). For example, the MFCCs, while appropriately scaling the frequency axis to account for perceptual resolution, do not take account of the wide perceptual difference between resonances and spectral valleys. First-order HMMs ignore the high degree of correlation across many frames of speech data (compensating by using delta coefficients is only a very rough use of speech dynamics). Intonation is widely ignored (e.g., F0) or treated as noise (e.g., durational factors), despite evidence of its use in human speech perception. Of course, in a practical world, you use whatever works, and current systems, despite their flaws, provide sufficiently high accuracy for small vocabularies (e.g., recognizing the digits in spoken telephone or credit card numbers, or controlling computer menu selections via voice). More widespread use of speech in telephone dialogs will await advances in both the quality of synthetic speech and the recognition accuracy of spontaneous conversations.

BIOGRAPHY

Douglas O'Shaughnessy has been a professor at INRS-Telecommunications at the University of Quebec and adjunct professor at McGill University in Montreal, Québec, Canada, since 1977. His interests include automatic speech synthesis, analysis, coding, and recognition. He has been an associate editor for the Journal of the Acoustical Society of America since 1998, and will be the general chair of the 2004 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in Montreal, Canada. Dr. O'Shaughnessy received all his degrees from the Massachusetts Institute of Technology (MIT), and is a fellow of the Acoustical Society of America.

He is the author of the textbook Speech Communications: Human and Machine (IEEE Press, 2000).

BIBLIOGRAPHY

- J. Deller, J. G. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- 2. P. Noll, Digital audio coding for visual communications, *Proc. IEEE* 83: 925–943 (1995).

- J. Johnston and K. Brandenburg, Wideband coding perceptual considerations for speech and music, in S. Furui and M. Sondhi, eds., Advances in Speech Signal Processing, Marcel Dekker, New York, 1992, pp. 109–140.
- 5. A. Gersho, Advances in speech and audio compression, *Proc. IEEE* 82: 900–918 (1994).
- D. O'Shaughnessy, Speech Communication: Human and Machine, Addison-Wesley, Reading, MA, 1987.
- A. Spanias, Speech coding: A tutorial review, *Proc. IEEE* 82: 1541–1582 (1994).
- P. Kroon and W. Kleijn, Linear predictive analysis by synthesis coding, in R. Ramachandran and R. Mammone, eds., Modern Methods of Speech Processing, Kluwer, Norwell, MA, 1995, pp. 51-74.
- R. Cox and P. Kroon, Low bit-rate speech coders for multimedia communication, *IEEE Commun. Mag.* 34(12): 34-41 (1996).
- W. Hess, Pitch and voicing determination, in S. Furui and M. Sondhi, eds., Advances in Speech Signal Processing, Marcel Dekker, New York, 1992, pp. 3–48.
- 11. L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- J. Piconi, Signal modeling techniques in speech recognition, *Proc. IEEE* 81: 1215–1247 (1993).
- J.-C. Junqua and J.-P. Haton, Robustness in Automatic Speech Recognition, Kluwer, Norwell, MA, 1996.
- 14. Y. Ephraim, Statistical-model-based speech enhancement systems, *Proc. IEEE* 80(10): 1526-1555 (1992).
- N. Morgan and H. Bourlard, Continuous speech recognition, IEEE Signal Process. Mag. 12(3): 25-42 (1995).
- D. Reynolds and R. Rose, Robust text-independent speaker indentification using Gaussian mixture speaker models, *IEEE Trans. SAP* 3: 72–83 (1995).
- D. Klatt, Review of text-to-speech conversion for English, J. Acoust. Soc. Am. 82: 737-793 (1987).
- J. van Santen, Using statistics in text-to-speech system construction, Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis, 1994, pp. 240-243.
- T. Dutoit, From Text to Speech: A Concatenative Approach, Kluwer, Norwell, MA, 1997.
- E. Moulines and F. Charpentier, Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Commun. 9: 453–467 (1990).

SPEECH RECOGNITION

JAYADEV BILLA BBN Technologies Cambridge, Massachusetts

1. INTRODUCTION

Speech is a natural and preferred form of communication among humans. Automatic speech recognition attempts to

extend this mode of communication to human-machine interaction. As a field of study, automatic speech recognition has a history extending back to the late 1960s. Over these years our understanding of the process of speech communication has increased substantially. This greater understanding, coupled with an even more developed ability to harness powerful computational models to encapsulate our knowledge of speech communication, has allowed for the development of increasingly capable speech recognition systems. Today, in some specialized applications, state-of-the-art speech recognition systems are capable of recognition and transcription of speech at performance levels that rival human ability.

In this article, we introduce the problem of automatic speech recognition and describe the reasons that make speech recognition by machine difficult. We also describe the science behind automatic speech recognition and discuss the principles and architecture of a typical large vocabulary speech recognition system. We then provide an overview of current state-of-the-art speech recognition research and applications, and conclude with a description of current trends in speech recognition system research and development.

2. WHY SPEECH RECOGNITION IS DIFFICULT

The ease with which humans use speech understates the complexity of this task for machines. The fundamental difficulty with speech recognition is the overwhelming variability in the production of speech. Indeed, if everyone always spoke in exactly the same way so that each word, whether spoken by the same person or different people, always sounded exactly the same, then we could simply store all the words and recognize speech by a simple comparison with the words we had heard and stored earlier. However, this is not the case; there is a considerable amount of variation in spoken speech even when the same speaker repeats the same sentence. Some forms of variability include those arising in the speaker due to mood, stress, and state of health, and those arising from the environment such as background noise or room acoustics. Other causes of variability in spoken speech include those due to gender, age, accent, or speaking rate among speakers and due to language differences, such as regional dialects, or formal versus conversational speaking style.

Figure 1 shows the speech signal and spectrogram for the utterance "gray whales". Spectrograms are pictures of the distribution of energy in frequency over time. The dark horizontal bars correspond to resonances of the vocal tract, and change their vertical frequency positions over time, depending on the sound being produced as well as on neighboring sounds. The dependency of each sound on its neighboring sounds further increases the number of ways in which a particular word can be pronounced. If this were not enough, normal human speech is a continuous stream of words without any interword separation, unlike the clearly separated words of written text. As an example of the difficulty in word boundary identification, consider spoken instances of the phrases "recognize speech" and "wreck a nice beach."



Figure 1. A spoken instance of the utterance "gray whales." The speech signal (bottom) is displayed as a simple amplitude plot over time. The spectrogram (top) shows the energy distribution of the utterance over frequency and time. The horizontal dark areas correspond to the resonances of the vocal tract.

An automatic speech recognition system must successfully address these variabilities in speech before any useful output can be provided to the user. Current understanding of the speech production process and the semantic and grammatic structure of language provides a means to model some of the variability. We are, however, limited by an incomplete knowledge of the process of speech communication. We do not really know how humans organize acoustic information and store knowledge of language; not do we know how cognitive processes are formed and used to perform recognition tasks. Current automatic speech recognition systems reflect these shortcomings and attempt to fit mathematical models to account for our incomplete understanding of the speech recognition problem. Tremendous advances in speech recognition technology have come from our ability to build sophisticated models that compensate for this lack of knowledge. Nevertheless, the lack of complete understanding of the speech communication process is an additional obstacle for automatic speech recognition systems to overcome, in addition to the inherent difficulty of the speech recognition task.

Moreover, automatic speech recognition systems must work effectively over the majority of acoustic environments that the systems may encounter. The difficulty in producing robust speech recognition systems that can work effectively over different acoustic environments is the primary reason for the limited use of such systems. Advances since 1980 have enabled us to build and deploy speech recognition systems for tasks ranging from responses to interactive voice response (IVR) systems with prompts of the type, "please say or press one to select choice one for ..., say or press two" to highly accurate transcription of dictation in certain specialized areas. The corresponding research recognition systems are close to achieving real-time transcription of general English such as in news broadcasts. Figure 2 charts commercial speech recognition system complexity, measured by speaking mode and vocabulary size, and its corresponding application area over time, against a projection for the near future. Early commercial speech recognition systems possessed vocabularies of several words that were spoken in isolation. More recently deployed speech recognition systems have had vocabularies in the tens of thousands of words and are capable of recognizing fluent speech. In the future, commercial systems are likely to recognize normal conversational speech with vocabulary sizes approaching that of humans.

3. THE SCIENCE BEHIND AUTOMATIC SPEECH RECOGNITION

The term *recognition* in the context of speech recognition is typically framed as a classification problem. Classification is when one has a finite set of possible outcomes, such as in a multiple-choice question, and attempts to classify an object or event of interest as one of these possibilities based on some observations. In automatic speech recognition systems, classification is usually considered a statistical pattern recognition problem. In the context of speech recognition, this involves building mathematical models of spoken speech and using them to identify the most likely sequence of words in the vocabulary as the recognized sentence or phrase.

A typical speech recognition system, illustrated in Fig. 3, consists of three main components: a *feature* extraction stage, where one extracts a set of speech features that can minimize some of the variability in speech without loss of information; a *training stage*, where one builds a set of mathematical models of speech; and a *recognition* stage, where one uses the trained models to make the classification of the spoken speech into a word or sentence.



Figure 2. Timeline of the increase in speech recognition system complexity and its corresponding area of application. Early systems were capable of recognition of a few words spoken in isolation. Future systems are likely to have vocabularies sizes similar to human vocabularies and capable of recognition of natural spoken language.



Figure 3. Components of a typical speech recognition system. A speech recognition system must first be trained (bottom) to generate models of speech. Recognition proceeds by comparing the input speech to the speech models to determine which speech model is closest to the input speech.

The catalog of mathematical models with which we compare features of a spoken word or sentence must be created beforehand, and often involves *levels of knowledge* akin to the many ways in which humans use prior knowledge to recognize speech. Typically, automatic speech recognition systems use at least two levels of knowledge: an *acoustic model* describing the actual acoustics of the speech signal, and a *language model* describing what word follows or is most likely to follow the current word or set of words.

The acoustics of a spoken sentence can be decomposed into a sequence of progressively finer acoustic events such as words, syllables, allophones, and phonemes. Fig. 4 shows the decomposition of the utterance "gray whales" into allophones and phonemes. The term phoneme has its origins in linguistics and is defined as the smallest unit of speech that can distinguish one word from another, such as *bit* versus *pit*. An allophone, on the other hand, is one of two or more variants of the same phoneme such as the phoneme $p \setminus in pin$ and spin. Each of these acoustic events can be modeled individually, and recognition can be achieved by identifying the appropriate sequence of acoustic events. The fundamental acoustic event is often determined by the application and by the amount of training data. For example, if we intend to build a recognizer that recognizes the 10 digits, 0-9, we can simply build a model for each of these digits. It is likely, in this case, that we will have many instances of each digit spoken by many people with which to build a representative model. On the other hand, if we intend to

Words	Gray			Whales			
Phonemes	-g	r	ey	w	ey	I	Z-
Allophones	- [g] r	g [r] ey	r [ey] w	ey [w] ey	w [ey] l	ey [l] z	l [z] -

Figure 4. Decomposition of the utterance "gray whales" into words, phonemes, and allophones. Depending on the application and availability of training data, any of these acoustic events can be modeled individually. Recognition is then achieved by identifying the appropriate sequence of acoustic events. build a speech recognition system for general dictation with a vocabulary of 60,000 words, it is unlikely that we will have many spoken instances for most of the words in the vocabulary. In these cases, it is advisable to build phonetic models as the number of phonemes is much smaller than the number of words and therefore far more likely to have sufficient training instances. These models of acoustic events encapsulate the characteristics of the actual spoken speech signal and form the acoustic model.

Acoustic models are created by organizing labeled samples of speech, specifically, spoken speech and its corresponding text, using a training paradigm. Most often, acoustic modeling is attempted via statistical models that are capable of automatically modeling the key distinguishing features of their training data. This ability to "learn" the key features from highly variable input is crucial to the application of statistical models to the speech recognition problem. Of these statistical models, hidden Markov models (HMMs) are the most widely used methodology for acoustic modeling. HMMs possess many interesting properties and are supported by elegant algorithms that allow for their application to the speech recognition problem.

A language model, in contrast, determines valid linguistic constructs and encapsulates the ways in which words are connected to form phrases and sentences. An alternate description of the language model is that it specifies the grammar that the speech recognizer must adhere to when producing a recognition result. Depending on the intended area of application of the speech recognition system, language modeling can be one of two types: rule-based or statistical. Rule-based language models are used when there is a rigid structure to the spoken sentence, such as a credit card number spoken by a person, and are created by writing explicit rules that specify allowable sentences. Statistical language models are used when there is no rigid structure to the spoken utterances, such as in normal conversational speech. Unlike rule-based language models, statistical language models do not explicitly specify all allowable sentences, but assign scores to each sequence or group of words, proportional to the frequency that a particular sequence or group of words occurs in the training data.

3.1. Feature Extraction

The goal of feature extraction is to translate the raw speech waveform into a set of features that effectively capture the salient aspects of the speech signal and at the same time reduce the effect of variability due to speaker or environment in the final representation. There is a considerable amount of information in the speech signal. Some information, such as the resonances of the vocal tract that correspond to vowel-like sounds in spoken words, is crucial to speech recognition. Other information, such as gender and mood of the speaker or the acoustic environment, provides no useful information for speech recognition.

Information in a speech signal is encapsulated mostly in the energy and in the frequency distribution of that energy. The spectral characteristics of speech change rapidly over time as different words are spoken. To capture these rapid changes in speech over the course of a word or sentence, features must be generated sequentially over the duration of the utterance. Furthermore, features must be generated on sufficiently short time segments so that the spectral characteristics are relatively invariant. Typically, small Fourier analysis algorithms such as the fast Fourier transform (FFT) are used to perform the spectral analysis of the speech segment. Since the analysis is performed on relatively short segments of speech, this mode of analyzing speech is often referred to as *short-time Fourier analysis*.

The source-filter model of speech production provides a convenient parametric model for feature extraction. Parametric modeling reduces the representation of a signal to the estimation of model parameters that, in turn, are likely to form a much more compact representation of the signal. Figure 5 provides a graphical view of a model of human speech production and its source-filter model equivalent. The idea behind the source-filter model is that speech is a result of airflow (source) being shaped by the



Figure 5. A model of human speech production and its corresponding source-filter model equivalent: (a) human speech production; (b) speech production model.

vocal tract (filter). The type of source excitation controls the realization of different types of speech. For example, voiced speech, which includes sounds with strong periodic structure such as vowels, can be generated with periodic source excitation. Unvoiced speech, which encompasses sounds with little or no periodic structure such as the \s in *set*, is generated with white-noise excitation. Further, it is assumed that information is carried primarily by the shape of the vocal tract rather than in the airflow, and that variability in airflow is the primary source of variability among speakers. These simplifying assumptions render the feature extraction problem mathematically tractable and allow the immediate applicability of many techniques from signal processing theory.

Figure 6 illustrates a typical speech feature extraction paradigm. From basic Fourier analysis, given a sourcefilter model, one can consider the speech spectrum to be the product of the excitation spectrum and the filter spectrum. Furthermore, if one takes the logarithm of the speech spectrum and transforms it back to the time domain (by way of the inverse Fourier transform) one can represent the excitation and vocal tract separately. The result of this inverse transform of the log spectrum is the *cepstrum*. The term *cepstrum* is actually derived from "spectrum" spelled with the first four letters in reverse order as *c-e-p-s*-trum to signify the inversion of the spectrum. Excitation and vocal tract have very different spectral characteristics; excitation is usually located in the higher order terms and the vocal tract in the lower-order terms of the cepstrum. The noninformative excitation can be easily discarded by considering only the lower-order cepstral terms.

This basic paradigm can be further enhanced by incorporating information from psychoacoustic studies. For example, the human auditory system is more sensitive to lower-frequency components than higher-frequency components. The differential emphasis of frequency content can be incorporated into the feature extraction paradigm by warping the frequency scale to compress the higher frequencies relative to the lower frequencies. This allows us to place more emphasis on low-frequency components over high-frequency components. Common



Figure 6. Block diagram of a typical speech feature extraction paradigm. Speech is first transformed into the spectral domain with an FFT algorithm. Various techniques, such as log compression and frequency warping, are used to emphasize or deemphasize characteristics of the speech signal depending on their importance for speech recognition. Finally, the modified spectrum is inverted using an inverse FFT algorithm to yield cepstra. The lower order ceptral terms typically form the speech features used during speech recognition.

warping strategies include mel-scale warping and barkscale warping, both of which are approximately linear to about 1000 Hz and logarithmic thereafter.

Excellent overviews of the various techniques for speech feature extraction can be found in the literature [1,2].

3.2. Model Training

Speech model training allows us to encapsulate the acoustics of the speech signal and the manner in which words are connected to form phrases and sentences. As mentioned earlier, the acoustics of the speech signal are captured in the acoustic model, and the language and word structure in the language model.

3.2.1. Acoustic Modeling. Acoustic modeling in current speech recognition systems is most commonly based on *hidden Markov models* (HMMs), statistical models that are capable of automatically extracting statistically significant information from available speech data. In HMM theory, the distance or closeness to incoming speech is measured in terms of the probability that the input speech could be generated by that model. This probabilistic approach allows us to absorb the variation in speech features and provides a powerful paradigm for speech recognition.

Often in physical processes, the current state of the process has a significant influence on subsequent events. This concept is embodied in the Markov property, which states that given the current state of the process or system, the future evolution of the system is independent of its past. Models based on the Markov property are called Markov models. Figure 7a shows a simple fourstate Markov model of weekly weather. Each state in this Markov model is associated with an output symbol indicative of a possible weather observation: hot, cold, windy, or cloudy. These weather observations are called output symbols because the Markov model can be regarded as a generative model; it outputs symbols as a transition is made from one state to the other. The arrows connecting the states indicate allowable transitions, and each number on the arcs corresponds to the transition probability: the probability of an arc being taken. For example, the self-arc on the "cloudy" state has a probability of 0.6, indicating that if one makes the observation that it was "cloudy" this week, then there will be a 60% chance of making the observation that the weather is still "cloudy" the following week. Note that in the Markov model, the transition from one state to another is probabilistic, but the production of the output symbol is deterministic and known. Also note that at any given instance the current state completely determines the set of all possible states to transition to without regard to how one transitioned into the current state. For example, in the Markov model of Fig. 7a, if the current weekly weather corresponds to the state "hot," next week's weather can only be hot, windy, or cold and this holds true irrespective of whether last week was observed as cloudy, hot, or cold.

As it turns out, this definition of Markov models is too restrictive to be useful in many interesting problems. In this example, it is too restraining to confine the output symbols from any state to correspond to a particular



Figure 7. A model of weekly weather patterns. A Markov model of weather is presented in (**a**). A corresponding hidden Markov model (HMM) is presented in (**b**). Note the output probability distributions in the HMM compared to the deterministic single output in the Markov model.

weather observation. Indeed, it is far more meaningful to describe the weekly weather as some combination of the four weather patterns: hot, cloudy, cold, or windy. In *hidden* Markov models (HMMs) the output of each state is associated with an output probability distribution rather than a deterministic and known outcome. In an HMM all output symbols are possible at each state, albeit with differing probabilities. The probabilities associated with each state are known as output probabilities.

Figure 7b shows the Markov model of Fig. 7a extended to the HMM formulation. The difference is that we now have a probability distribution associated with each state, and when a transition is made into a state, the output symbol is chosen according this probability distribution. Since weather of any type is possible with varying likelihoods in every state, given a sequence of weekly weather readings, the state sequence is unobservable and *hidden*.

The HMM methodology is very well suited for speech recognition and provides a natural and highly reliable way of modeling and recognizing speech. In the case of speech, we can visualize the states as corresponding to functional (or largely unchanging) portions of the word or subword, such as the beginning or end, and the speech features as corresponding to the outputs of the state. HMMs simultaneously model time and feature variability. Time variability is modeled by state-transition probabilities, and feature variability by state output probability distributions. Training the HMM involves estimation of the output state probability distributions and the state-transition probabilities. Figure 8 shows the structure of a typical three-state speech HMM. The structure of a speech HMM has a natural flow of transitions from left to right to indicate the forward time flow of speech. In a typical speech recognition system there is one HMM per phonetic context, to reflect the differing pronunciation of phonemes in different contexts. Although different phonetic contexts could have different structures, the HMM structure is usually kept the same, with HMMs being distinguished from each other by their transition and output probabilities.

Thus far we have referred to neither the language or type of acoustic event that the HMM is modeling. This brings out an important quality of statistical modeling in general, and HMMs in particular: that the HMM is independent of language and type of acoustic event. An HMM does not require any specific changes for language or type of acoustic event.

An HMM can be regarded as a generative model. For example, in Fig. 8, as we enter into state 1, a speech feature is emitted. Based on the transition probabilities out of state 1, a transition can be made back to state 1, 2, or 3, and another speech feature emitted based on the probability distribution corresponding to the state to which the transition was made. This continues until a transition is made out of state 3. At that point, the sequence of emitted speech features corresponds to the phoneme or phonetic context for which the HMM had been constructed.

In a recognition paradigm, the same HMM can be used to solve the reverse problem — given a sequence of speech features, what is the probability that the HMM could have generated this speech feature sequence? In this mode, given a starting speech feature, one can estimate



the probability of the feature being generated by the probability distribution of state 1. One can now assume that a transition is made from state 1 to state 2 and estimate the probability that the next speech feature was generated by the probability distribution of state 2. This is continued until the last state is exited. The product of all the encountered output probabilities and transition probabilities gives the probability that this specific state sequence generated the speech feature sequence. For every possible sequence of states, one obtains a different probability value. During recognition, the probability computation is performed for all speech HMMs and all possible state sequences. The state sequence and speech HMM that gives the highest probability is declared to be the recognized phoneme or phonetic context.

Mathematically tractable techniques such as the Baum–Welch reestimation and Viterbi algorithms provide elegant automatic solutions to these problems and have allowed HMMs to be successful in a wide variety of difficult speech recognition applications. In-depth presentations of HMMs and their application to speech recognition can be found in Rabiner and Juang [2] and Huang et al. [3].

3.2.2. Language Modeling. A language model acts as a grammar dictating which word can or cannot follow another word or group of words. Without a language model, every word in the vocabulary of the speech recognition system is equally likely and must be considered at the end of every other word. Language modeling provides a means to limit the vocabulary at each decision point either by eliminating unlikely words or by increasing the probability of more likely words.

There are two approaches to language modeling for speech recognition: rule-based and statistical. Rule-based language modeling involves the construction of a set of rules that encompass all allowable sentences. Since the set of rules is explicit and must specify all allowable sentences, rule-based language models are primarily used in very restrictive tasks such as IVR systems. A sample rulebased language model for a speech recognition system, used to recognize names for automatic telephone dialing by name, is shown in Fig. 9. Allowable names are John Doe, Jane Smith, Simon Smith, and Mike Phillips with an additional garbage name. Since a rule-based grammar completely specifies all the allowable output of the speech recognizer, a valid name is always recognized even when invalid names are spoken. The catchall garbage word is often included to prevent the recognition of one of the allowable names when an invalid input is presented to



Figure 8. Typical structure of a three-state speech HMM. Numbers on the arcs indicate transition probabilities and p_1 , p_2 , and p_3 refer to state output probability distributions. HMM training involves the estimation of both the transition and state output probabilities.

Figure 9. A rule-based language model for recognizing names for an automatic dialing task. The "garbage" name is used to prevent incorrect results with invalid input.

the system. Since rule-based language modeling imposes a rigid structure on the recognized utterance, it provides an excellent means for task automation. In the namedialer example, the person using the system would be required to say the first and last name in that order. Any valid name can be dialed out simply by looking up the name in the directory. The advantage of rigid structure is also the greatest shortcoming of rule-based language modeling. To describe a speech recognition task of any complexity, it becomes increasingly harder to describe all valid sentences. Consider, for example, designing a rulebased grammar to recognize an interview on television. It would be virtually impossible to describe the myriad ways in which the conversation can even start, let alone proceed.

Statistical language models, on the other hand, are based on probability estimates of the likelihood that one word will follow one or more other words. Because they are estimated directly from training data, statistical models have the advantage of be able to encapsulate the colloquial characteristics of speech in addition to its semantic and syntactic elements. However, statistical language modeling requires a large pool of training data. Without a large training data set, many word sequences are either not observed or are observed far too infrequently to reliably estimate probabilities. Despite this constraint, statistical language models have contributed significantly to improvement in speech recognition system performance, especially in large vocabulary speech recognition tasks, where it is far too complex and virtually impossible to build rule-based language models. A detailed discourse on statistical language modeling can be found Jelinek's treatise [4].

4. CURRENT STATE OF SPEECH RECOGNITION AND FUTURE DIRECTIONS

Automatic speech recognition technology has now advanced to the point where speech recognition systems with vocabularies of a hundred to several thousand words can be robustly deployed in a variety of tasks, ranging from service automation to general consumer applications to the management of spoken information. Table 1 shows the error rates of current high-performance speech recognition systems on various tasks. Observe the increase in error rates as the vocabulary of the speech recognition system and the complexity of the speech increases.

A logical area of deployment has been in the telephony market, where speech is the primary modality for information exchange. Telephony applications include service automation, an early example of which is the automation of operator services with *yes/no* responses in the collect call prompt: "You have a collect call from ..., to accept this call press one or say yes". A telephony application rapidly now gaining popularity is voice dialing. Today, many vendors offer the ability to dial numbers via simple voice commands such as "call home."

Consumer applications of speech recognition have been primarily in the dictation and computer command and control arenas, allowing users to dictate for a variety of word processing tasks or to perform basic computer commands such as "open file" or "close file." Some popular products aimed for the general consumer include Dragon Systems' Dragon NaturallySpeaking, IBM's ViaVoice, and L&H's Voice Xpress.

A relatively new area of application of speech recognition technology has been in the management of spoken information. The conversion of spoken information into textual form allows rapid access to information which would otherwise require the arduous task of listening to large archives of spoken material. An example of an application providing this ability is BBN's Rough 'n' Ready (RnR) audio indexer system [5]. The RnR system is capable of real-time speech recognition of broadcast news, and provides information extraction and indexing coupled with a Web-enabled interface allowing for rapid retrieval of spoken information in a computer accessible form.

Although speech recognition by machine is increasingly used in popular applications, significant issues remain. The translation of speech recognition technology from the laboratory into the "real world" is still fraught with problems. Fielded systems typically show higher error rates than systems evaluated in the laboratory, due mostly to the lack of control over the working environment. To address this vulnerability, current speech recognition systems are normally trained on large amounts of realworld training data. This ensures that, when fielded outside the laboratory, the speech recognition system can handle the unconstrained environment of the real world. Typical real-world training data include recordings of broadcast news such as from CNN or NPR (National Public Radio), and recordings, taken with participants' consent, of telephone conversations between friends and family. Broadcast news recordings include the accompanying commercials, background music, and noise. On the other hand, telephone conversation between family and friends is typical of normal conversational speech with the attendant disfluencies. State-of-the-art speech recognition systems have an average word error rate of close to 10% on broadcast speech, and close to 30% on conversational speech, with vocabularies of up to 65,000 words [6,7].

Table 1. Error Rates of High-Performance Speech Recognition Systems in 2001

Task	Type of Speech (complexity)	Vocabulary Size (Words)	Word Error Rate (%)
Connected digits	Read	10	< 0.3
Automated travel agent	Spontaneous	2,500	< 2
Broadcast news	Read and spontaneous (mixed)	64,000	15
Telephone conversations	Spontaneous and conversational	45,000	35

2390 SPEECH RECOGNITION

While the use of real-world training data has enhanced the usefulness of speech technology in everyday applications, significant barriers remain to the seamless integration of speech technology in day-to-day environments. Some issues currently being addressed in speech recognition research laboratories to improve the practical use of speech technology are presented below.

4.1. Speaker and Environmental Adaptation and Robustness

Current speech recognition is often handicapped by the fact that it typically does not run "out of the box." Noise in the environment is a common reason for the failure of speech recognition systems. In some cases, even using a different microphone to capture the speech signal can significantly change the properties of the resultant speech features, with consequent loss in system performance [8]. Speaker and environment adaptation allow an existing speech recognition system to handle new speakers and environment without substantial change in performance. Adaptation can be performed in two different paradigms: supervised, where the system is adapted based on enrollment data, or unsupervised, where the system adapts automatically without any specific input to the system beyond that spoken for the purpose of recognition.

4.2. Conversational Speech

In normal spontaneous spoken conversation, there is a significant amount of disfluency such as hesitations and false starts. In addition, conversational speech commonly suffers from the lack of clear articulation as well as other speaker maladies such as highly variable speaking rate or increased emotional emphasis. These factors combine to cause significant problems for accurate speech recognition. It would not be unusual to see a doubling in error rates by moving from broadcast news to conversational speech recognition task with all other factors remaining the same.

4.3. Speed Versus Accuracy

Another issue with state-of-art speech recognition systems is the large amount of time needed to provide the most accurate recognition result or transcript. Speech recognition systems often need to be tuned to run in real time by sacrificing performance. This is slowly changing due to the ever-increasing speed of computers as well as due to algorithmic improvements. Nevertheless, significant work still needs to be done to allow for real-time speech recognition without losing accuracy.

4.4. Language and Task Independence

Most speech recognition systems are built for specific languages and, more often than not, specific tasks within those languages. This artifact of the specificity of the training that goes into the system invariably requires retraining the system if the language or task changes. The idea behind language and task independence is to lessen the effect of specificity by either generalization of the system or by developing the system's capability to automatically acquire new language or task skills.

4.5. Understanding

After speech recognition, understanding speech is the next frontier in human-machine interaction. Give a system capable of consistently accurate speech recognition, the next logical step would be to use this ability to extract the meaning of the spoken words and provide the appropriate response.

5. CONCLUSION

Speech recognition is perhaps one of a handful of technologies that have the potential to fundamentally change the way we interact with machines. The science of speech recognition continues to develop, and as it does, so will its impact on our lives. Even now, simple applications such as voice dialing for mobile phones or voice controllable car radios are making everyday tasks both safe and convenient. In the future, these and other emerging applications will surely change the nature of human-machine interaction and be responsible for many new and exciting changes to the ways we work and live.

BIOGRAPHY

Jayadev Billa received the B.E. degree in electronics and communication engineering in 1991 from Osmania University, Hyderabad, India, and M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pennsylvania, in 1993 and 1997, respectively. He joined the BBN Technologies, Cambridge, Massachusetts, in 1996 where he is now a senior scientist. At BBN he works on the design and development of large vocabulary speech recognition system in a variety of languages. His areas of interest are speech feature extraction and acoustic modeling for speech recognition.

BIBLIOGRAPHY

- 1. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.
- L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models* for Speech Recognition, Edinburgh Univ. Press, Edinburgh, Scotland, 1990.
- 4. F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1998.
- 5. J. Makhoul et al., Speech and language technologies for audio indexing and retrieval, *Proc. IEEE* **88**(8): 1338–1354 (Aug. 2000).
- S. Matsoukas et al., The 1998 BBN Byblos primary system applied to English and Spanish broadcast news transcription, *Proc. DARPA Broadcast News Transcription and Understand*ing Workshop, Herndon, UK, March 1999.
- 7. J. Billa et al., Recent experiments in large vocabulary conversational speech recognition, in *Proc. Int. Conf. Acoustics*, *Speech and Signal Processing*, Phoenix, AZ, IEEE, April 1999.
- A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, Boston, 1993.

SPREAD SPECTRUM SIGNALS FOR DIGITAL COMMUNICATIONS

JOHN G. PROAKIS Northeastern University Boston, Massachusetts

1. INTRODUCTION

Spread spectrum signals is a class of signals that were designed primarily for use in digital communication systems to overcome either intentional or unintentional interference. Such signals were originally used in military communication systems either to provide resistance to jamming or to hide the signal by transmitting it at low power and, thus, making it difficult for an unintended listener to detect its presence in noise (low probability of intercept). However, today, spread spectrum signals are used to provide reliable communications in a variety of commercial applications. For example, spread spectrum signals are used in the so-called unlicensed frequency bands, such as the Industry, Scientific, and Medical (ISM) band at 2.4 GHz. Typical applications in the ISM band are cordless telephones, wireless LANs, and cable replacement systems such as Bluetooth. Since the band is unlicensed, there is no central control over the radio resources, and the systems have to function even in the presence of severe interference from other communication systems and other electrical and electronic equipment (e.g., microwave ovens, radars, etc.). Here the interference is not intentional, but the interference may nevertheless be enough to disrupt the communication for non-spread spectrum systems.

Code-division multiple access systems (CDMA systems) use spread spectrum techniques to provide communication to several concurrent users. CDMA is used in one second generation (IS-95) and several third generation wireless cellular systems (e.g., cdma2000 and WCDMA). One advantage of using interference-resistant signals in these applications is that the radio resource management (primarily the channel allocation to the active users) is significantly reduced.

The name spread spectrum stems from the fact that the transmitted signals occupy a much wider frequency band than is required to transmit the information. There are many different ways to spread the bandwidth of the information-bearing signal. The most common ones are called direct-sequence (DS) and frequency-hopping (FH) spread spectrum (SS). In DS-SS, the information-bearing signal is spread over the entire channel bandwidth in a manner that appears random. In a FH-SS, the transmitter changes the carrier frequency of the relatively narrowband information-bearing signal in a fashion that appears random. At any given time, only a small fraction of the available channel bandwidth is used and exactly which fraction used is known only to the intended receiver. A hybrid of DS and FH spread spectrum is also possible. Several other methods are available for spreading the bandwidth of the information-bearing signals; however, the clear majority of the implemented systems are either DS or FH or a hybrid of DS/FH.

The literature in the field of spread spectrum communications is quite voluminous and ranges from text books to specialized conference and journal papers. For an interesting review of the history of the development of spread spectrum, we recommend the papers [1-3]. Among the available books, we would like to especially mention the text by Simon, Omura, Scholtz, and Levitt [4] which covers quite a lot of the classical spread spectrum techniques. The books by Ziemer, Peterson, and Borth [5] and by Dixon [6], which cover more of the current commercial applications, are also recommended. The reference lists of the above mentioned books contain several thousand entries. Among the tutorial-style papers that are available in the literature, we would like to especially mention the 1982 paper by Pickholtz, Schilling, and Milstein [7].

2. MODEL OF A SPREAD SPECTRUM DIGITAL COMMUNICATION SYSTEM

The basic elements of a spread spectrum digital communication system are illustrated in Fig. 1. We observe that the channel encoder and decoder and the modulator and demodulator are the basic elements of a conventional digital communication system. In addition to these elements, a spread-spectrum system employs two identical pseudorandom sequence generators, one which interfaces with the modulator at the transmitting end and the second which interfaces with the demodulator at the receiving end. These two generators produce a pseudorandom or pseudonoise (PN) binary-valued sequence, which is used to spread the transmitted signal at the modulator and to despread the received signal at the demodulator.

Time synchronization of the PN sequence generated at the receiver with the PN sequence contained in the received signal is required to properly despread the received spread-spectrum signal. In a practical system, synchronization is established prior to the transmission



Figure 1. Model of spread-spectrum digital communications system.

of information by transmitting a fixed PN bit pattern which is designed so that the receiver will detect it with high probability in the presence of interference. After time synchronization of the PN sequence generators is established, the transmission of information commences. In the data mode, the communication system usually tracks the timing of the incoming received signal and keeps the PN sequence generator in synchronism. The synchronization of spread spectrum signals is treated in Refs. 4 and 7 and in the article by Luise et al. [8] in this encyclopedia.

Interference is introduced in the transmission of the spread-spectrum signal through the channel. The characteristics of the interference depend to a large extend on its origin. The interference may be generally categorized as being either broadband or narrowband (partial band) relative to the bandwidth of the informationbearing signal, and either continuous in time or pulsed (discontinuous) in time. For example, an interfering signal may consist of a high-power sinusoid in the bandwidth occupied by the information-bearing signal. Such a signal is narrowband. As a second example, the interference generated by other users in a multipleaccess channel depends on the type of spread-spectrum signals that are employed by the various users to transmit their information. If all users employ broadband signals. the interference may be characterized as an equivalent broadband noise. If the users employ frequency hopping to generate spread-spectrum signals, the interference from other users may be characterized as narrowband or partial band.

Our discussion will focus on the performance of spreadspectrum signals for digital communication in the presence of narrowband and broadband interference. Two types of digital modulation are considered, namely, PSK and FSK. PSK modulation is appropriate for applications where phase coherence between the transmitted signal and the received signal can be maintained over a time interval that spans several symbol (or bit) intervals. This is usually the case in DS-SS, where a single carrier frequency is modulated by the spread spectrum signal that covers the entire channel bandwidth. On the other hand, FSK modulation with noncoherent detection is appropriate in applications where phase coherence of the carrier cannot be accurately estimated. This is usually the case in FH-SS, where the carrier frequency is hopped rapidly, typically at the transmitted symbol (or bit) rate. In such a case, the relatively short time interval spanned by a single symbol (or bit) is not sufficient to obtain an accurate estimate of the carrier phase.

The PN sequence generated at the modulator is used in conjunction with the PSK modulation to shift the phase of the PSK signal pseudorandomly, as described below at a rate that is an integer multiple of the bit rate. The resulting modulated signal is called a *direct-sequence* (DS) spread-spectrum signal. When used in conjunction with binary or M-ary (M > 2) FSK, the PN sequence is used to select the carrier frequency of the transmitted signal pseudorandomly. The resulting signal is called a *frequency-hopped* (FH) spread-spectrum signal.

3. DIRECT-SEQUENCE SPREAD-SPECTRUM SYSTEMS

Let us consider the transmission of a binary information sequence by means of binary PSK. The information rate is R bits/sec and the bit interval is $T_b = 1/R$ sec. The available channel bandwidth is W Hz, where $W \gg R$. At the modulator, the bandwidth of the information signal is expanded to W Hz by shifting the phase of the carrier pseudorandomly at a rate of W times/sec according to the pattern of the PN generator. The basic method for accomplishing the spreading is shown in Fig. 2.

The information-bearing baseband signal is denoted as $\upsilon(t)$ and is expressed as

$$\upsilon(t) = \sum_{n=-\infty}^{\infty} a_n g_T(t - nT_b) \tag{1}$$

where $\{a_n = \pm 1, -\infty < n < \infty\}$ and $g_T(t)$ is a rectangular pulse of duration T_b . This signal is multiplied by the signal from the PN sequence generator, which may be expressed as

$$c(t) = \sum_{n=-\infty}^{\infty} c_n p(t - nT_c)$$
⁽²⁾

where $\{c_n\}$ represents the binary PN code sequence of ± 1 's and p(t) is a rectangular pulse of duration T_c , as illustrated in Fig. 2. This multiplication operation serves to spread the bandwidth of the information-bearing signal (whose bandwidth is R hz, approximately) into the wider bandwidth occupied by PN generator signal c(t)



Figure 2. Generation of a DS spread-spectrum signal.



Figure 3. Convolution of spectra of the (**a**) data signal with the (**b**) PN code signal.

(whose bandwidth is $1/T_c$, approximately). The spectrum spreading is illustrated in Fig. 3, which shows, in simple terms, using rectangular spectra, the convolution of the two spectra, the narrow spectrum corresponding to the information-bearing signal and the wide spectrum corresponding to the signal from the PN generator.

The product signal v(t)c(t), also illustrated in Fig. 2, is used to amplitude modulate the carrier $A_c \cos 2\pi f_c t$ and, thus, to generate the double-sideband, suppressed carrier (DSB-SC) signal

$$u(t) = A_c v(t)c(t)\cos 2\pi f_c t \tag{3}$$

Since $v(t)c(t) = \pm 1$ for any *t*, it follows that the carriermodulated transmitted signal may also be expressed as

$$u(t) = A_c \cos[2\pi f_c t + \theta(t)] \tag{4}$$

where $\theta(t) = 0$ when $\upsilon(t)c(t) = 1$ and $\theta(t) = \pi$ when $\upsilon(t)c(t) = -1$. Therefore, the transmitted signal is a binary PSK signal.

The rectangular pulse p(t) is usually called a *chip* and its time duration T_c is called the *chip interval*. The reciprocal $1/T_c$ is called the *chip rate* and corresponds (approximately) to the bandwidth W of the transmitted signal. The ratio of the bit interval T_b to the chip interval T_c usually is selected to be an integer in practical spread spectrum systems. We denote this ratio as

$$L_c = \frac{T_b}{T_c} \tag{5}$$

Hence, L_c is the number of chips of the PN code sequence/information bit. Another interpretation is that L_c



Figure 4. Demodulation of DS spread spectrum signal.

represents the number of possible 180° phase transitions in the transmitted signal during the bit interval T_b .

The demodulation of the signal is performed as illustrated in Fig. 4. The received signal is first multiplied by a replica of the waveform c(t) generated by the PN code sequence generator at the receiver, which is synchronized to the PN code in the received signal. This operation is called (spectrum) despreading, since the effect of multiplication by c(t) at the receiver is to undo the spreading operation at the transmitter. Thus, we have

$$A_c \upsilon(t) c^2(t) \cos 2\pi f_c t = A_c \upsilon(t) \cos 2\pi f_c t \tag{6}$$

since $c^2(t) = 1$ for all t. The resulting signal $A_c \upsilon(t) \cos 2\pi f_c t$ occupies a bandwidth (approximately) of R hz, which is the bandwidth of the information-bearing signal. Therefore, the demodulator for the despread signal is simply the conventional cross correlator or matched filter. Since the demodulator has a bandwidth that is identical to the bandwidth of the despread signal, the only additive noise that corrupts the signal at the demodulator is the noise that falls within the information-bandwidth of the received signal.

3.1. Effect of Despreading on a Narrowband Interference

It is interesting to investigate the effect of an interfering signal on the demodulation of the desired informationbearing signal. Suppose that the received signal is

$$r(t) = A_c \upsilon(t)c(t)\cos 2\pi f_c t + i(t) \tag{7}$$

where i(t) denotes the interference. The despreading operation at the receiver yields

$$r(t)c(t) = A_c \upsilon(t) \cos 2\pi f_c t + i(t)c(t) \tag{8}$$

The effect of multiplying the interference i(t) with c(t), is to spread the bandwidth of i(t) to W HZ.

As an example, let us consider a sinusoidal interfering signal of the form

$$i(t) = A_I \cos 2\pi f_I t \tag{9}$$

where f_I is a frequency within the bandwidth of the transmitted signal. Its multiplication with c(t) results in a wideband interference with power-spectral density $I_0 = P_I/W$, where $P_I = A_I^2/2$ is the average power of the interference. Since the desired signal is demodulated by a matched filter (or correlator) that has a bandwidth R,

the total power in the interference at the output of the demodulator is

$$I_0 R_b = P_I R_b / W = \frac{P_I}{W/R_b} = \frac{P_I}{T_b / T_c} = \frac{P_I}{L_c}$$
 (10)

Therefore, the power in the interfering signal is reduced by an amount equal to the bandwidth expansion factor W/R. The factor $W/R = T_b/T_c = L_c$ is called the *processing* gain of the spread-spectrum system. The reduction in interference power is the basic reason for using spreadspectrum signals to transmit digital information over channels with interference.

In summary, the PN code sequence is used at the transmitter to spread the information-bearing signal into a wide bandwidth for transmission over the channel. By multiplying the received signal with a synchronized replica of the PN code signal, the desired signal is despread back to a narrow bandwidth while any interference signals are spread over a wide bandwidth. The net effect is a reduction in the interference power by the factor W/R, which is the processing gain of the spread-spectrum system.

The PN code sequence $\{c_n\}$ is assumed to be known only to the intended receiver. Any other receiver that does not have knowledge of the PN code sequence cannot demodulate the signal. Consequently, the use of a PN code sequence provides a degree of privacy (or security) that is not possible to achieve with conventional modulation. The primary cost for this security and performance gain against interference is an increase in channel bandwidth utilization and in the complexity of the communication system.

3.2. Probability of Error

In this section we derive the probability of error for a DS spread spectrum system assuming that the information is transmitted via binary PSK. Within the bit interval $0 \le t \le T_b$, the transmitted signal is

$$s(t) = a_0 g_T(t) c(t) \cos 2\pi f_c t, \quad 0 \le t \le T_b$$

$$(11)$$

where $a_0 = \pm 1$ is the information symbol, the pulse $g_T(t)$ is defined as

$$g_T(t) = \begin{cases} \sqrt{\frac{2\mathcal{E}_b}{T_b}}, & 0 \le 1 \le T_b \\ 0, & \text{otherwise} \end{cases}$$
(12)

and c(t) is the output of the PN code generator which, over a bit interval, is expressed as

$$c(t) = \sum_{n=0}^{L_c - 1} c_n p(t - nT_c)$$
(13)

where L_c is the number of chips per bit, T_c is the chip interval, and $\{c_n\}$ denotes the PN code sequence. The PN code chip sequence $\{c_n\}$ is designed to be uncorrelated (white); that is,

$$E(c_n c_m) = E(c_n)E(c_m) \text{ for } n \neq m$$
(14)

and each chip is +1 or -1 with equal probability. These conditions imply that $E(c_n) = 0$ and $E(c_n^2) = 1$.

The received signal is assumed to be corrupted by an additive interfering signal i(t). Hence,

$$r(t) = a_0 g_T(t) c(t) \cos(2\pi f_c t + \phi) + i(t)$$
(15)

where ϕ represents the carrier phase shift. Since the received signal r(t) is typically the output of an ideal bandpass filter in the front end of the receiver, the interference i(t) is also a bandpass signal, and may be represented as

$$i(t) = i_c(t) \cos 2\pi f_c t - i_s(t) \sin 2\pi f_c t$$
(16)

where $i_c(t)$ and $i_s(t)$ are the two quadrature components of i(t).

We assume that the receiver is perfectly synchronized to the received signal and the carrier phase is perfectly estimated by a PLL. Then, the signal r(t) is demodulated by first despreading through multiplication by c(t) and then crosscorrelation with $g_T(t)\cos(2\pi f_c t + \phi)$, as shown in Fig. 5. At the sampling instant $t = T_b$, the output of the correlator is

$$y(T_b) = \mathcal{E}_b + y_i(T_b) \tag{17}$$

where $y_i(T_b)$ represents the interference component, which has the form

$$y_{i}(T_{b}) = \int_{0}^{T_{b}} c(t)i(t)g_{T}(t)\cos(2\pi f_{c}t + \phi) dt$$
$$= \sum_{n=0}^{L_{c}-1} c_{n} \int_{0}^{T_{b}} p(t - nT_{c})i(t)g_{T}(t)\cos(2\pi f_{c}t + \phi) dt$$
$$= \sqrt{\frac{2\mathcal{E}_{b}}{T_{b}}} \sum_{n=0}^{L_{c}-1} c_{n}v_{n}$$
(18)

where, by definition,

$$v_n = \int_{nT_c}^{(n+1)T_c} i(t) \cos(2\pi f_c t + \phi) \, dt \tag{19}$$

The probability of error depends on the statistical characteristics of the interference component. Its mean value is

$$E[y_i(T_b)] = 0 \tag{20}$$



Figure 5. DS spread-spectrum signal demodulator.

Its variance is

$$E[y_i^2(T_b)] = \frac{2\mathcal{E}_b}{T_b} \sum_{n=0}^{Lc-1} \sum_{m=0}^{Lc-1} E(c_n c_m) E(v_n v_m)$$

But $E(c_n c_m) = \delta_{mn}$. Therefore,

$$E[y_i^2(Tb)] = \frac{2\mathcal{E}_b}{T_b} \sum_{n=0}^{L_c-1} E(v_n^2)$$
$$= \frac{2\mathcal{E}_b}{T_b} L_c E(v^2)$$
(21)

where $v = v_n$, as given by Eq. (9).

Consider for sinusoidal interfering signal at the carrier frequency, that is,

$$i(t) = \sqrt{2P_I} \cos(2\pi f_c t + \Theta_I) \tag{22}$$

where P_I is the average power and Θ_I is the phase of the interference, which is random and uniformly distributed over the interval $(0, 2\pi)$. If we substitute for i(t) in Eq. (19), it is easy to show that

$$E(v^2) = \frac{T_c^2 P_I}{4}$$
(23)

and, therefore,

$$E[y_i^2(T_b)] = \frac{\mathcal{E}_b P_I T_c}{2} \tag{24}$$

The ratio of $\{E[y(T_b)]\}^2$ to $E[y_i^2(T_b)]$ is the SNR at the detector. In this case we have

$$(\text{SNR})_D = \frac{\mathcal{E}_b^2}{\mathcal{E}_b P_I T_c/2} = \frac{2\mathcal{E}_b}{P_I T_c}$$
(25)

To see the effect of the spread-spectrum signal, we express the transmitted energy \mathcal{E}_b as

$$\mathcal{E}_b = P_S T_b \tag{26}$$

where P_S is the average signal power. Then, if we substitute for \mathcal{E}_b in Eq. (25) we obtain

$$(\text{SNR})_D = \frac{2P_S T_b}{P_I T_c} = \frac{2P_S}{P_I/L_c}$$
(27)

where $L_c = T_b/T_c$ is the processing gain. Therefore, the spread-spectrum signal has reduced the power of the interference by the factor L_c .

Another interpretation of the effect of the spreadspectrum signal on the sinusoidal interference is obtained if we express $P_I T_c$ in Eq. (27) as follows. Since $T_c \simeq 1/W$, we have

$$P_I T_c = P_I / W = I_0 \tag{28}$$

where I_0 is the power-spectral density of an equivalent interference in a bandwidth W. Therefore, in effect, the spread-spectrum signal has spread the sinusoidal interference over the wide bandwidth W, creating an equivalent spectrally flat noise with power-spectral density I_0 . Hence,

$$(\text{SNR})_D = \frac{2\mathcal{E}_b}{I_0} \tag{29}$$

The probability of error for a DS spread-spectrum system with binary PSK modulation is easily obtained from the SNR at the detector, if we make an assumption on the probability distribution of the sample $y_i(T_b)$. From Eq. (18) we note that $y_i(T_b)$ consists of a sum of L_c uncorrelated random variables $\{c_n v_n, 0 \le n \le L_c - 1\}$, all of which are identically distributed. Since the processing gain L_c is usually large in any practical system, we may use the Central Limit Theorem to justify a Gaussian probability distribution for $y_i(T)$. Under this assumption, the probability of error is

$$P_b = Q\left(\sqrt{\frac{2\mathcal{E}_b}{I_0}}\right) \tag{30}$$

where I_0 is the power-spectral density of an equivalent broadband interference and Q(x) is defined as

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^{2}/2} dt$$
 (31)

Hence, the effect of the interference on the error probability is equivalent to that of broadband white Gaussian noise with spectral density I_0 .

3.3. The Interference Margin

We may express $rac{\mathcal{E}_b}{I_0}$ in the *Q*-function in Eq. (30) as

$$\frac{\mathcal{E}_b}{I_0} = \frac{P_S T_b}{P_I/W} = \frac{P_S/R}{P_I/W} = \frac{W/R}{P_I/P_S}$$
(32)

Also, suppose we specify a required E_b/I_0 to achieve a desired level of performance. Then, using a logarithmic scale, we may express Eq. (32) as

$$10 \log \frac{P_I}{P_S} = 10 \log \frac{W}{R} - 10 \log \frac{\mathcal{E}_b}{I_0}$$
$$\left(\frac{P_I}{P_S}\right)_{\rm dB} = \left(\frac{W}{R}\right)_{\rm dB} - \left(\frac{\mathcal{E}_b}{I_0}\right)_{\rm dB}$$
(33)

The ratio $(P_I/P_S)_{dB}$ is called the *interference margin*. This is the relative power advantage than an interference may have without disrupting the communication system.

For example, suppose we require an $(\mathcal{E}_b/I_0))_{\rm dB} = 10 \text{ dB}$ to achieve reliable communication. What is the processing gain that is necessary to provide an interference margin of 20 dB? Clearly, if W/R = 1000, then $(W/R)_{\rm dB} = 30$ dB and the interference margin is $(P_I/P_S)_{\rm dB} = 20$ dB. This means that the average interference power at the receiver may be 100 times the power P_S of the desired signal and we can still maintain reliable communication.

3.4. Performance of Coded Spread-Spectrum Signals

It is shown in many textbooks on digital communications (for reference, see [9], Chapter 8), that when the transmitted information is coded by a binary linear (block or convolutional) code, the SNR at the output of a softdecision decoder in the presence of spectrally flat Gaussian interference is increased by the coding gain, defined as

$$\operatorname{coding \,gain} = R_c d_{\min}^H \tag{34}$$

where R_c is the code rate and d_{\min}^H is the minimum Hamming distance of the code. Therefore, the effect of coding is to increase the interference margin by the coding gain. Thus, Eq. (33) may be modified as

$$\left(\frac{P_I}{P_S}\right)_{\rm dB} = \left(\frac{W}{R}\right)_{\rm dB} + (CG)_{\rm dB} - \left(\frac{\mathcal{E}_b}{I_0}\right)_{\rm dB} \tag{35}$$

where $(CG)_{dB}$ denotes the coding gain. Typical coding gains obtained by use of binary block or convolutional codes are in the range of 4 to 7 dB.

3.5. Pulsed Interference

A very damaging type of interference for DS-SS is broadband pulsed noise, whose power is spread over the entire system bandwidth W. The pulsed interference is transmitted for a fraction ρ of the time, that is, ρ is the duty cycle of the transmitted interference, where $0 < \rho \leq 1$. If this signal is being transmitted by a jammer, this allows the jammer to transmit pulses with a power level P_I/ρ for ρ percent of the time, with an equivalent spectral density of $I_0 = P_I/W$, where P_I is the average transmitted power. For simplicity, let us assume that the interference pulse spans an integer number of symbols (or bits) and that the pulsed noise is Gaussian distributed. When the interferer is not transmitting, the received information bits are assumed to be error-free, and when the interferer is transmitting, the probability of error for an uncoded DS-SS system is

$$P(\rho) = \rho Q\left(\sqrt{\frac{2\mathcal{E}_b}{I_0}\rho}\right) \tag{36}$$

The worst case duty cycle that maximizes the probability of error for the communication system can be found by differentiating $P(\rho)$ with respect to ρ . Thus, we find that the worst-case pulsed noise occurs when

$$\rho^* = \begin{cases} \frac{0.71}{\mathcal{E}_b/I_0}, & \frac{\mathcal{E}_b}{I_0} \ge 0.71\\ 1, & \frac{\mathcal{E}_b}{I_0} < 0.71 \end{cases}$$
(37)

and the corresponding probability of error is

$$P(\rho^*) = \begin{cases} \frac{0.082}{\mathcal{E}_b/I_0} = \frac{0.082P_I/P_S}{W/R}, & \frac{\mathcal{E}_b}{I_0} \ge 0.71\\ Q\left(\sqrt{\frac{2\mathcal{E}_b}{I_0}}\right) = Q\left(\sqrt{\frac{2W/R}{P_I/P_S}}\right), & \frac{\mathcal{E}_b}{I_0} < 0.71 \end{cases}$$
(38)

The error rate performance given by Eq. (38) for $\rho = 1.0, 0.1, 0.01$, and 0.001 along with the worst-case performance based on ρ^* is plotted in Fig. 6. When we compare the error rate for continuous wideband Gaussian noise interference ($\rho = 1$) with worst-case pulse interference, we find a large difference in performance; for example, approximately 40 dB at an error rate of 10^{-6} . This is, indeed, a large penalty.

If we simply add coding to the DS spread-spectrum system, the performance in SNR is improved by an amount equal to the coding gain, which in most cases is limited to less than 10 dB. The reason that the addition of coding does not improve the performance significantly is that the interfering signal pulse duration (duty cycle) may be selected to affect many consecutive coded bits. Consequently, the code word error probability is high due to the burst characteristics of the interference.

In order to improve the performance of the coded DS spread-spectrum system, we should interleave the coded bits prior to transmission over the channel. The effect of interleaving is to make the coded bits that are affected by the interferer statistically independent. Figure 7 illustrates a block diagram of a DS spread-spectrum system that employs coding and interleaving. By selecting a sufficiently long interleaver so that the burst characteristics of the interference are eliminated, the penalty in performance due to pulse interference is significantly reduced; for example, to the range of 3–5 dB for conventional binary block or convolutional codes (for reference, see [9], Chapter 13).

4. FREQUENCY-HOPPED SPREAD SPECTRUM

In frequency-hopped (FH) spread spectrum, the available channel bandwidth W is subdivided into a large number of nonoverlapping frequency slots. In any signaling interval the transmitted signal occupies one or more of the available frequency slots. The selection of the frequency slot (*s*) in each signal interval is made pseudorandomly according to the output from a PN generator.

A block diagram of the transmitter and receiver for an FH spread-spectrum system is shown in Fig. 8. The modulation is either binary or M-ary FSK (MFSK). For example, if binary FSK is employed, the modulator selects one of two frequencies say f_0 or f_1 , corresponding to the transmission of a 0 for a 1. The resulting binary FSK signal is translated in frequency by an amount that is determined by the output sequence from a PN generator, which is used to select a frequency f_c that is synthesized by the frequency synthesizer. This frequency is mixed with the output of the FSK modulator and the resultant frequency-translated signal is transmitted over the channel. For example, by taking *m* bits from the PN generator, we may specify $2^m - 1$ possible carrier frequencies. Figure 9 illustrates an FH signal pattern.

At the receiver, there is an identical PN sequences generator synchronized with the received signal, which is used to control the output of the frequency synthesizer. Thus, the pseudorandom frequency translation introduced at the transmitter is removed at the demodulator by mixing the synthesizer output with the received signal.



Figure 7. Black diagram of communication system with coding and interleaving.

The resultant signal is then demodulated by means of an FSK demodulator. A signal for maintaining synchronism of the PN sequence generator with the FH received signal is usually extracted from the received signal.

Although binary PSK modulation generally yields better performance than binary FSK, it is difficult to maintain phase coherence in the synthesis of the frequencies used in the hopping pattern and, also, in the propagation of the signal over the channel as the signal is hopped from one frequency to the another over a wide bandwidth. Consequently, FSK modulation with noncoherent demodulation is usually employed in FH spread-spectrum systems.

The frequency-hopping rate, denoted as R_h , may be selected to be either equal to the symbol rate, or lower than the symbol rate, or higher than the symbol rate. If R_h is equal to or lower than the symbol rate, the FH system is called a slow-hopping system. If R_h is higher than

Figure 6. Bit error probability for BPSK modulated DS spread spectrum with pulsed interference having a duty cycle P.

the symbol rate; that is, there are multiple hops/symbols, the FH system is called a fast-hopping system. However, there is a penalty incurred in subdividing an information symbol into several frequency-hopped elements, because the energy from these separate elements is combined noncoherently (for reference, see [9], Chapter 12).

FH spread-spectrum signals may be used in CDMA where many users share a common bandwidth. In some cases, an FH signal is preferred over a DS spread-spectrum signal because of the stringent synchronization requirements inherent in DS spread-spectrum signals. Specifically, in a DS system, timing and synchronization must be established to within a fraction of a chip interval $T_c = 1/W$. Conversely, in an FH system, the chip interval T_c is the time spend in transmitting a signal in a particular frequency slot of bandwidth $B \ll W$. But this interval is approximately 1/B, which is much larger than 1/W. Hence, the timing requirements in an FH system.

4.1. Slow Frequency-Hopping Systems

40

Let us consider a slow frequency-hopping system in which the hop rate $R_h = 1$ hop/bit. If the interference on the channel is broadband and is characterized as AWGN with power-spectral density I_0 , the probability of error for the detection of noncoherently demodulated binary FSK is

$$P_b = \frac{1}{2} e^{-\mathcal{E}_b / 2I_0}$$
(39)

where \mathcal{E}_b/I_0 is the SNR/bit.

As in the case of a DS spread-spectrum system, we observe that \mathcal{E}_b , the energy/bit, can be expressed as $\mathcal{E}_b = P_S T_b = P_S/R$, where P_S is the average transmitted power and R is the bit rate. Similarly, $I_0 = P_I/W$, where



Figure 8. Black diagram of an FH spread-spectrum system.



Figure 9. An example of an FH pattern.

 P_I is the average power of the broadband interference and *W* is the available channel bandwidth. Therefore, the SNR/bit can be expressed as

$$\frac{\mathcal{E}_b}{I_0} = \frac{W/R}{P_I/P_S} \tag{40}$$

where W/R is the processing gain and P_I/P_S is the interference margin for the FH spread-spectrum signal.

Slow FH spread-spectrum systems are particularly vulnerable to partial-band interference that may result in FH CDMA systems. To be specific, suppose that the partial-band interference is modeled as a zero-mean Gaussian random process with a flat power-spectral density over a fraction of the total bandwidth W and zero in the remainder of the frequency band. In the region or regions where the power-spectral density is nonzero, its value is I_0/ρ , where $0 < \rho < 1$. In other words, the interference average power P_I is assumed to be constant.

Let us consider the worst-case partial-band interference by selecting the value of ρ that maximizes the error

probability. In an uncoded slow-hopping system with binary FSK modulation and noncoherent detection, the transmitted frequencies are selected with uniform probability in the frequency band W. Consequently, the received signal will be corrupted by interference with probability ρ . When the interference is present, the probability of error is $\frac{1}{2} \exp(-\rho \mathcal{E}_b/2I_0)$ and when it is not, the detection of the signal is assumed to be error free. Therefore, the average probability of error is

$$P_b(\rho) = \frac{\rho}{2} e^{-\rho \mathcal{E}_b/2I_0}$$
$$= \frac{\rho}{2} \exp\left(\frac{\rho W/R}{2P_I/P_S}\right)$$
(41)

Figure 10 illustrates the error rate as a function of \mathcal{E}_b/I_0 for several values of ρ . By differentiating $P_b(\rho)$, and solving for the value of ρ that maximizes $P_b(\rho)$, we find

$$\rho^* = \begin{cases} 2I_0/\mathcal{E}_b, & \rho \ge 2\\ 1, & \mathcal{E}_b/I_0 < 2 \end{cases}$$
(42)

The corresponding error probability for the worst case partial-band interference is

$$P_{b} = \begin{cases} e^{-1}/\mathcal{E}_{b}/I_{0}, & \mathcal{E}_{b}/I_{0} \ge 2\\ \frac{1}{2}e^{-\mathcal{E}_{b}/2I_{0}}, & \mathcal{E}_{b}/I_{0} < 2 \end{cases}$$
(43)

which is also shown in Fig. 10. Whereas the error probability decreases exponentially for full-band interference as given by Eq. (39), the error probability for worst-case partial band interference decreases only inversely with \mathcal{E}_b/I_0 . This result is similar to the error probability for DS spread-spectrum signals in the presence of pulse interference. It is also similar to the error probability for binary PSK in a Rayleigh fading channel.

An effective method for combatting partial band interference in a FH system is signal diversity, which can be obtained by simple repetition of the transmitted information bit on different frequencies (or by means of block or convolutional coding). Signal diversity obtained through coding provides a significant improvement in performance relative to uncoded signal transmission. In fact, it has been shown by Viterbi and Jacobs [10] that by optimizing the code design for the partial-band



Figure 10. Performance of binary FSK with Partial-band interference.

interference, the communication system can achieve an average bit-error probability of

$$P_b = e^{-\mathcal{E}_b/4I_0} \tag{44}$$

Therefore, the probability of error achieved with the optimum code design decreases exponentially with an increase in SNR and is within 3 dB of the performance obtained in an AWGN channel. Thus, the penalty due to partial-band interference is reduced significantly.

4.2. Fast Frequency-Hopping Systems

In fast FH systems, the frequency-hop rate R_h is some multiple of the symbol rate. Basically, each (M-ary) symbol interval is subdivided into N subintervals, which are called *chips* and one of M frequencies is transmitted in each subinterval. Fast frequency-hopping systems are particularly attractive for military communications. In such systems, the hop rate R_h may be selected sufficiently high so that a potential intentional interferer does not have sufficient time to detect the presence of the transmitted frequency and to synthesize a jamming signal that occupies the same bandwidth.

To recover the information at the receiver, the received signal is first dehopped by mixing it with the hopped carrier frequency. This operation removes the hopping pattern and brings the received signal in all subintervals (chips) to a common frequency band that encompasses the M possible transmitted frequencies. The signal in each subinterval is then passed through the M matched filters (or correlators) tuned to the M possible transmitted frequencies which are sampled at the end of each subinterval and passed to the detector. The detection of the FSK signals is noncoherent. Hence, decisions are based on the magnitude of the matched filter (or correlator) outputs.

Since each symbol is transmitted over N chips, the decoding may be performed simply on the basis of hard decisions for the chips on each hop.

To determine the probability of error for the detector, we recall that the probability of error for noncoherent detection of binary FSK for each hop is

$$p = \frac{1}{2}e^{-\mathcal{E}_b/2NI_0} \tag{45}$$

where N is the number of frequency hops (chips) per bit, and \mathcal{E}_b/N is the energy of the signal per hop. Assuming that N is odd, the decoder decides in favor of the transmitted binary FSK bit that is larger in at least (N + 1)/2 chips. Thus, the decision is made on the basis of a majority vote given the decisions on the N chips. Consequently, the probability of a bit error is

$$P_b = \sum_{m=(N+1)/2}^{N} {\binom{N}{m}} p^m (1-p)^{N-m}$$
(46)

where p is given by Eq. (45). We should note that the error probability P_b for hard-decision decoding of the N chips will be higher than the error probability for a single hop/bit PSK system, which is given by Eq. (39), when the SNR/bit \mathcal{E}_b/I_0 is the same in the two systems.

5. COMMERCIAL SPREAD SPECTRUM SYSTEMS

As mentioned earlier, the number of nonmilitary spread spectrum systems have increased rapidly the last decades. The applications are quite diverse: underwater communications [11], wireless local loop systems [12], wireless local area networks, cellular systems, satellite communications [13], and ultra wideband systems [14]. Spread spectrum is also used in wired application in, for example, power-line communication [15] and have been proposed for communication over cable-TV networks [12] and optical fiber systems [16,17]. Finally, spread spectrum techniques have been found to be useful in ranging, such as, radar and navigation and the Global Positioning System (GPS) [18]. Other applications are watermarking of multimedia [19] and (mentioned here as a curiosity) in clocking of high-speed electronics [20,21]. Due to space constraint, we will only briefly mention some of the hot wireless applications here.

The wireless local area network (WLAN) standard IEEE 802.11 was originally designed to operate in the ISM band at approximately 2.4 GHz. The standard supports several different coding and modulation formats and several data rates. The first version of the standard was released in 1997 and supports both FH and DS spread spectrum formats with data rates of 1 or 2 Mbit/s [22]. The FH modes are slow hopping and use so-called Gaussian FSK (GFSK) modulation (binary for 1 Mbit/s and 4-ary for 2Mbit/s). The system hops over 79 subcarriers with 1-MHz spacing. The DS-SS modes use a 11-chip long Barker sequence which is periodically repeated for each symbol. The chip rate is 11 Mchips/s, and the symbol rate is 1 Msymbols/s. The modulation is differentially encoded BPSK or QPSK (for 1 and 2 Mbit/s, respectively). We note that the processing gain is rather low, especially for the DS-SS modes.

The 802.11 standard has since 1997 been extended in several directions (new bands, higher data rates, etc). In 1999, the standard was updated to IEEE 802.11b (also known as Wi-Fi, if the equipment, also passes an interoperability test). In addition to the original 1 and 2 Mbit/s modes, IEEE 802.11b also supports 5.5 and 11 Mbit/s DS-SS modes [23] and several other optional modes with varying rates. The higher rate DS-SS modes uses so-called complementary code keying (CCK). The chip rate is still 11 Mchips/s and each symbol is represented by 8 complex chips. Hence, for the 5.5-Mbit/s mode, each symbol carries 4 bits, and for the 11-Mbit/s mode, each symbol carries 8 bits. Hence, the processing gains is reduced compared to the 1- and 2-Mbit/s modes. As a matter of fact, the 11 Mbit/s is perhaps not even a spread-spectrum system. The CCK modulation is a little bit complicated to describe, but in essence it forms the complex chips by combining a block code and differential QPSK [22], Section 18.4.6.5].

Bluetooth is primarily a cable replacement system, that is, a system for short range communication with relatively low-data rate. It is designed for the ISM band and uses slow hopping FH-SS with GFSK modulation (BT = 0.5 and modulation index between 0.28 and0.35). The system hops over 79 subcarriers with a rate of 1600 hop/s. The subcarrier spacing is 1 MHz, and in most countries the subcarriers are placed at $f_k = 2402 + k$ MHz for $k = 0, 1, \dots, 78$. Bluetooth supports both synchronous and asynchronous links and several different coding and packet schemes. The user data rates varies from 64 kbits/s (symmetrical and synchronous) to 723 kbits/s (asymmetrical and asynchronous). The maximum symmetrical rate is 434 kbits/s. The range of the system is quite short, probably less than 10 m in most environments. It is likely that future versions of Bluetooth will support higher data rates and longer ranges.

The first cellular system with a distinct spread spectrum component was IS-95 (also known as cdmaOne or somewhat pretentiously as CDMA). Although the Global System for Mobile Communications (GSM), has a provision for frequency hopping, it is not usually considered to be a spread spectrum system. Often, spread spectrum and codedivision multiple access (CDMA) are used as synonyms, although they really are not. A multiple access method is a method for allowing several links (that are not at the same geographical location) to share a common communication resource. CDMA is a multiple access method where the links are spread spectrum links. A receiver that is tuned to a certain user relies on the anti-jamming properties of the spread spectrum format to suppress the other users' signals.

IS-95 uses DS-SS links with a chip rate of 1.2288 Mchips/s and a bandwidth of (approximately) 1.25 MHz. In the downlink (forward link or base-station-to-terminal link) the chips are formed by a combination of convolutional encoding, repetition encoding, and scrambling. The chips are transmitted both in inphase and quadrature (but scrambled by different PN sequences). In the uplink (reverse link), the transmitted chips are formed by a combination of convolutional coding, orthogonal block coding, repetition coding, and scrambling. In the original IS-95 (IS-95A), the uplink was designed such that the detection could be done noncoherently. In the third generation evolution of IS-95, known as cdma2000, the modulation and coding has changed and the transmitted bandwidth tripled to allow for peak data rates exceeding 2 Mbit/s [24, -26].

Another third generation system is Wideband CDMA (WCDMA) [27]. WCDMA is a rather complex system with many options and modes. We will here only briefly describe the frequency-division duplex (FDD) mode. The FDD mode uses direct-sequence spreading with a chip rate of 3.84 Mchips/s. The chip waveform is a root-raised cosine pulse with roll-off factor 0.22, and the bandwidth of the transmitted signal is approximately 5 MHz. WCDMA supports many different information bit rates by changing the spreading factor (from 4 to 256 in the uplink and from 4 to 512 in the downlink) and the error control scheme (no coding, convolutional coding, or turbo coding); however, the modulation is QPSK with coherent detection in all cases. Today, the maximum information data rate is roughly 2 Mbits/s, but it is likely that future revisions of the standard will support higher rates through new combinations of spreading, coding, and modulation.

Acknowledgment

The author wishes to thank Professors Erik G. Strom, Tony Ottosson, and Arne Svensson for contributing Section 5 of this article.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982-1984; Interim Dean of the College of Engineering, 1992-1993; Chairman of the Department of Electrical and Computer Engineering, 1984-1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering,
adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book Digital Communications (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books Introduction to Digital Signal Processing (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), Digital Signal Processing Laboratory (Prentice-Hall, Englewood Cliffs, NJ, 1991); Advanced Digital Signal Processing (Macmillan, New York, 1992), Algorithms for Statistical Signal Processing (Prentice-Hall, Englewood Cliffs, NJ, 2002), Discrete-Time Processing of Speech Signals (Macmillan, New York, 1992, IEEE Press, New York, 2000), Communication Systems Engineering (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), Digital Signal Processing Using MATLAB V.4 (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and Contemporary Communication Systems Using MATLAB (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

- 1. Robert A. Scholtz, The origins of spread-spectrum communications, *IEEE Trans. Commun.* **30**(5): 822–854 (May 1982). (Part 1).
- Robert A. Scholtz, Notes on spread-spectrum history, *IEEE Trans. Commun.* 31(1): 82–84 (Jan. 1983).
- R. Price, Further notes and anecdotes on spread-spectrum origins, *IEEE Trans. Commun.* 31(1): 85–97 (Jan. 1983).
- 4. Marvin K. Simon, Jim K. Omura, Robert A. Scholtz, and Barry K. Levitt, *Spread Spectrum Communications Handbook*, revised edition McGraw-Hill, 1994.
- Rodger E. Ziemer, Roger L. Peterson, and David E. Borth, Introduction to Spread Spectrum Communications, Prentice-Hall, 1995.
- 6. Robert C. Dixon, Spread Spectrum Systems with Commercial Applications, 3rd ed., John Wiley and Sons, 1994.
- R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications — A tutorial. *IEEE Trans. Commun.* **30**(5): 855–884 (May 1982).
- 8. M. Luise, U. Mengali, and M. Morelli, Synchronization in Digital Communications, *The Wiley Encyclopedia of Telecommunications*.
- 9. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
- A. J. Viterbi and I. M. Jacobs, Advances in Coding and Modulation for Noncoherent Channels Affected by Fading, Partial Band, and Multiple-Access Interference, in A. J. Viterbi ed., Advances in Communication Systems, Vol. 4, Academic, New York, 1975.

- Charalampos C. Tsimenidis, Oliver R. Hinton, Alan E. Adams, and Bayan S. Sharif, Underwater acoustic receiver employing direct-sequence spread spectrum and spatial diversity combining for shallow-water multiaccess networking, *IEEE J. Oceanic Engineering* 26(4): 594-603 (Oct. 2001).
- D. Thomas Magill, Spread spectrum techniques and applications in America, Proceedings of International Symposium on Spread Spectrum Techniques and Applications, 1–4, 1995.
- D. Thomas Magill, Francis D. Natali, and Gwyn P. Edwards, Spread-spectrum technology for commercial applications, *Proceedings of the IEEE* 82(4): 572-584 (April 1994).
- Moe Z. Win and Robert A. Scholtz, Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communications, *IEEE Trans. Commun.* 48(4): 679-691 (April 2000).
- Denny Radford, Spread spectrum data leap through AC power wiring, *IEEE Spectrum* 33(11): 48–53 (Nov. 1996).
- Jawad A. Salehi, Code division multiple-access techniques in optical fiber networks — Part I: fundamental principles, *IEEE Trans. Commun.* 37(8): 824–833 (Aug. 1989).
- Jawad A. Salehi and Charles A. Brackett, Code division multiple-access techniques in optical fiber networks—Part II: Systems performance analysis. *IEEE Trans. Commun.* 37(8): 834-842 (Aug. 1989).
- Michael S. Braasch and A. J. van Dierendonck, GPS receiver architectures and measurements, *Proceedings of the IEEE* 87(1): 48–64 (Jan. 1999).
- Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon, Secure spread spectrum watermarking for multimedia, *IEEE Trans. Image Processing* 6(12): 1673–1687 (Dec. 1997).
- Harry G. Skinner and Kevin P. Slattery, Why spread spectrum clocking of computing devices is not cheating, *Proceedings 2001 International Symposium on Electromagnetic Compatibility* 537-540 (Aug. 2001).
- S. Gardiner, K. Hardin, J. Fessler, and K. Hall, An introduction to spread-spectrum clock generation for EMI reduction, *Electronic Engineering* **71**(867): 75, 77, 79, 81 (April 1999).
- 22. IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements-part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. IEEE Std 802.11-1997, November 1997. ISBN: 1-55937-935-9.
- 23. Supplement to IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements- part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Higher-speed physical layer extension in the 2.4 GHz band. IEEE Std 802.11b-1999, January 2000. ISBN: 0-7381-1811-7.
- Daisuke Terasawa and Jr. Edward G. Tiedemann, cdmaOne (IS-95) technology overview and evolution. In Proceedings IEEE Radio Frequency Integrated Circuits (RFIC) Symposium, 213-216, June 1999.
- Douglas N. Knisely, Sarath Kumar, Subhasis Laha, and Sanjiv Nanda, Evolution of wireless data services: IS-95 to cdma2000, *IEEE Commun. Mag.* 36(10): 140-149 (Oct. 1998).
- Theodore S. Rappaport, Wireless Communications: Principles and Practice, 2nd ed., Prentice-Hall, Upper Saddle River, NJ 2001.

2402 STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE

27. 3rd generation partnership project; technical specification group radio access network; physical layer - general description (release 4), 3GPPTS 25.201 V4.1.0 (2001-12). Online: http://www.3gpp.org. Accessed March 20, 2002.

STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE*

ARTHUR A. GIORDANO AG Consulting Inc., LLC Burlington, Massachusetts

THOMAS A. SCHONHOFF Titan Systems Corporation Northboro, Massachusetts

1. INTRODUCTION

A statistical characterization of impulsive noise is a multifaceted task. For the characterization to be meaningful and useful in predicting communications performance, knowledge of the underlying physical mechanisms is required. In order to develop a statistical noise model that accurately represents the communications channel under investigation, physical measurements of the impulsive noise characteristics are generally needed. With this understanding, a statistical noise model can be developed and used to construct optimum or near-optimum detectors, estimate physical system and noise parameters, and determine communications performance. As a consequence, there is no single noise model that can be used as a representation of all impulsive noise channels. This article describes a number of important noise models where the reader is cautioned to first confirm that the appropriate noise model is selected for the communications channel under investigation.

Gaussian noise, which is the predominant noise source associated with thermal noise in receiver front-ends and numerous communications channels such as microwave and satellite channels, is analytically tractable and used to predict communications performance for a broad class of communications systems. Noise statistics are characterized by a Gaussian probability density function (pdf) where only knowledge of the mean and variance of the noise is needed to completely characterize the noise statistics. The spectral characteristics of the noise may be white or nonwhite, with both cases having been extensively investigated [23]. Perhaps the most common and simple case is associated with additive white Gaussian noise (AWGN) where a number of analytical performance results have been obtained [33]. In fact, Gaussian noise is so prevalent that all other cases are categorized as non-Gaussian. This categorization is perhaps unfortunate in that it leads to the assumption that non-Gaussian noise has a single representation.

Impulse noise is typically associated with noise pulses that have large peak amplitudes and bandwidths that generally exceed the receiver bandwidth. The tails of the pdf of impulse noise are greater in extent than that of Gaussian noise and often lead to moments of the distribution that are ill defined. In contradistinction to the Gaussian case where only the pdf is needed, statistical characterization of an impulsive noise process requires multiple statistics for a complete characterization. The most extensively studied statistic is acknowledged to be the amplitude probability distribution (APD) and is defined as the probability that the noise envelope exceeds a specified value. This statistic is referred to as a first-order statistic. A complete characterization of the noise process requires higher-order statistics generally associated with the time statistics of the noise. Examples of these statistics include the autocorrelation, the pulse spacing distribution (also termed pulse interarrival times), the pulse duration distribution, and the average envelope crossing rate.

When a Gaussian noise model is accepted as a meaningful model for the communications channel under investigation, it is well known [48] that the optimum detector is linear and is implemented as either a matched filter or a correlation receiver.¹ A consequence of this result is that bit error rate (BER) computations are often analytically tractable for a large class of modulations. In contrast, for an impulsive noise process the optimum detector is nonlinear with a structure that is dependent on the noise model utilized. Thus, many different receiver structures can be derived where each structure is optimized for the specified noise model.

To determine the detector structure for an impulsive noise model, two approaches, referred to as analytical and ad hoc, are followed. The analytical approach is based on an accurate model of the noise statistics developed from the physical processes that generate the noise. The ad hoc approach is based on suboptimal detector structures that utilize nonlinearities that are easily implemented and reduce the tails in the noise distribution resulting in improved performance over linear detectors operating in the same impulsive noise environment. This latter approach has the advantage that the receiver performance is likely to be less sensitive to the noise model and therefore may offer more robust performance in time-varying or unknown noise statistics.

2. CHARACTERIZATION OF IMPULSIVE NOISE

Common examples of communication channels with impulsive noise include atmospheric radio noise, man-made noise, telephone communications, underwater acoustic noise, and magnetic recording noise. Atmospheric radio noise, arising from lightning discharges in the atmosphere, is an electromagnetic interference that can seriously degrade receiver performance [43,44]. Man-made noise occurs in automotive ignitions [22], electrical machinery such as welders, power transmission and distribution lines [32], medical and scientific apparatus, and so on. In telephone communications impulsive noise is generated

^{*} This article is adapted, with permission, from a chapter in a textbook to be published by Prentice-Hall.

 $^{^1}$ In the non–white noise case, the detector uses a whitening filter prior to detection.

from telephone switches and is particularly important in characterizing the performance of high-speed digital subscriber loops [20,21,34,46]. These examples represent a diverse subset of impulsive noise environments where knowledge of the physical characteristics dictates the noise model selection.

To maintain generality but provide a basis for relating the noise model to a physical case, the important case of atmospheric radio noise will be emphasized. Although much of the treatment that follows is specific to this channel, the process is likely to be extendable to other physical channels.

A summary of several principal noise models is provided.

2.1. Statistical-Physical Model

The noise $\mathbf{n}(t)$ is represented over an interval T by a summation of impulses filtered by the channel and expressed as

$$\mathbf{n}(t) = \sum_{i=1}^{N} \mathbf{a}_i \delta(t - \mathbf{t}_i)$$
(1)

where **N** is the number of impulses occurring in the interval T, \mathbf{a}_i is a random variable (rv) representing the strength of the *i*th impulse and \mathbf{t}_i is the occurrence time of the *i*th impulse.

N is typically assumed to be Poisson distributed with parameter μT where μ is the average rate of arrival of the impulses. Other investigators postulate a Poisson-Poisson distribution [8] or a Pareto [24] distribution leading to different pulse clustering behavior than that of a Poisson. For example, in the Poisson-Poisson case clusters of noise pulses occur where the pulses within a cluster occur at a Poisson rate μ_p and the clusters themselves are Poisson at a slower rate than μ_p .

The rv \mathbf{a}_i can be often assumed to be Gaussian [34]. However, for the atmospheric radio noise case, the statistical physical model is well founded based on measurements and analytical modeling of the physical behavior of the communications environment. Giordano has shown that the rv \mathbf{a}_i is proportional to the received field strength which depends on the source to receiver distance \mathbf{r}_i in accordance with a generalized propagation law g, where $\mathbf{a}_i = g(\mathbf{r}_i)$ [9]. By assuming a spatial distribution of noise sources and a specific propagation law, the pdf of the rv \mathbf{a}_i can be determined.

Middleton has extended the statistical-physical model developed in Giordano by introducing more general assumptions on the noise spatial conditions and propagation assumptions [25,26,28]. These models are termed "canonical" in that their form is invariant of the physical source mechanisms. Two important cases, referred to as Class A and Class B, were developed. Class A models are used for the narrowband case where the noise bandwidth is comparable or less than the receiver front-end bandwidth and Class B models are used in the wideband case.²

A Class C model has also been defined consisting of a combination of Class A and B models.

2.2. Generalized-t or Hall Model

This model, developed by Hall [13], has the form represented by a the product of a zero mean, narrowband Gaussian process $\mathbf{n}_g(t)$ with variance σ_1^2 and a slowly varying modulating stationary process $\mathbf{a}(t)$ that is independent of $\mathbf{n}_g(t)$, that is

$$\mathbf{n}(t) = \mathbf{a}(t)\mathbf{n}_g(t) \tag{2}$$

This model is premised on the fact that the noise sources vary with time over a large dynamic range and that, unlike a Gaussian noise source where energy is delivered at a constant rate, the impulses tend to occur in bursts. To model atmospheric radio noise, the slowly varying modulating process $\mathbf{a}(t)$ is selected to behave in a way that is similar to that of an empirical model obtained from measurements.

A variant of this model referred to as a truncated Hall model [35] removes some of the undesirable attributes of the Hall model where moments associated with the APD are undefined. This model is obtained by truncating the pdf of the noise envelope thereby producing finite moments and a noise process that is physically realizable. It has a further advantage in that the impulsiveness of the noise can be specified by means of the parameter Vd, which is the rms to average envelope ratio.

2.3. Mixture Model

The mixture model [29] is obtained by judiciously selecting two noise models which can be added together to produce the desired noise statistics. Thus, the term mixture model actually represents a collection of models since the two subsidiary noise processes can be chosen from many possibilities. One form of the mixture model is represented as a zero mean noise process $\mathbf{n}_0(t)$ that is,

$$\mathbf{n}_0(t) = (1 - \varepsilon_r)\mathbf{n}_g(t) + \varepsilon_r \mathbf{n}_i(t) \tag{3}$$

where $\mathbf{n}_g(t)$ is a Gaussian noise process, $\mathbf{n}_i(t)$ is an impulsive noise process and ε_r is a constant with $0 < \varepsilon_r < 1$. (An alternate formulation can be obtained by adding a combination of noise envelopes.) This model is appealing in that for small amplitudes approaching zero, $\mathbf{n}_0(t)$ approaches a Gaussian distribution ($\varepsilon_r \rightarrow 0$); for large noise amplitudes which create "spikes" of noise, the heavy tails of the impulsive noise distribution predominate ($\varepsilon_r \rightarrow 1$).

2.4. Empirical Model

The empirical model is based on graphical fits of the APD [7]. This model assumes that the small amplitude part of the APD is represented by a Rayleigh distribution and that the large amplitude region can be represented by a log-normal or power Rayleigh distribution. The parameters for the distributions are based on measurements of three statistical moments which are the average noise envelope, the rms noise

 $^{^{2}}$ That is, the noise bandwidth is wider than the receiver frontend bandwidth.

2404 STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE

envelope, and the average logarithm of the noise envelope. Denoting the noise envelope as v(t), the average noise envelope V_{ave} over an interval T is given by

$$V_{\rm ave} = 1/T \int_0^T v(t) \, dt \tag{4}$$

Similarly, the rms noise envelope $V_{\rm rms}$ and average logarithm of the noise envelope $V_{\rm log}$ are given respectively by

$$V_{\rm rms} = \left[1/T \int_0^T v^2(t) \, dt \right]^{1/2} \tag{5}$$

and

$$V_{\rm log} = {\rm antilog}\left(1/T \int_0^T \log v(t) \, dt\right) \tag{6}$$

These moments are ordinarily measured in one bandwidth and are converted to other bandwidths as needed [39].

3. APD COMPUTATION USING HANKEL TRANSFORMS

When the statistical-physical model given in Eq. (1) is adopted, a very powerful method based on Hankel transforms is available to compute the APD and relate it to the underlying physical environment [9]. This method has been utilized in the atmospheric noise channel but is not restricted to that case. The method in fact applies to any narrowband receiver output representation when the input is a time sequence of independent impulses. An outline of the approach is presented here with a more complete derivation provided in Ref. 9.

Let us assume that a narrowband receiver with a carrier frequency f_c has an impulse response given by

$$h(t) = \beta(t)\cos(\omega_c t - \psi)u(t) \tag{7}$$

where $\beta(t)$ is the amplitude of h(t), ψ is the phase, $\omega_c = 2\pi f_c$, and u(t) is the unit step function. If the noise process given by Eq. (1) is applied to the input of a receiver with an impulse response given by Eq. (7), the output noise process $\mathbf{n}_0(t)$ can be expressed as

$$\mathbf{n}_0(t) = \sum_{i=1}^{N} \mathbf{v}_i(t) \cos(\omega_c t - \omega_c \mathbf{t}_i - \psi) u(t - \mathbf{t}_i)$$
(8)

where $\mathbf{v}_i(t) = \mathbf{a}_i \beta(t - \mathbf{t}_i)$. The output noise can also be expressed in terms of its envelope $\mathbf{v}(t)$ and phase $\psi_s(t)$ as

$$\mathbf{n}_0(t) = \mathbf{v}(t)\cos(\omega_c t + \psi_s(t)) \tag{9}$$

The **N** terms in Eq. (8) can now be viewed as **N** vectors (or phasors) where the *i*th vector amplitude is $\mathbf{v}_i(t) = \mathbf{a}_i \beta(t - \mathbf{t}_i)$ with a phase $\psi_i = -\omega_c \mathbf{t}_i - \psi$ as shown in Fig. 1. The vectors then sum to the total vector with amplitude $\mathbf{v}(t)$ and phase $\psi_s(t)$.

The Hankel transform is the tool required to relate $\mathbf{v}(t)$ to $\mathbf{v}_i(t)$ for a fixed value of $\mathbf{N} = k$. The method is based on the characteristics function (cf) theorem for determining the pdf of a sum of independent random vectors where the cf of the sum vector is the product of the cf's of the individual vectors. It is now assumed



Figure 1. Phasor representation of output noise process.

that the phase of each vector ψ_i is uniformly distributed. Then for a sum of independent random vectors with uniformly distributed phase, the Hankel transform of the magnitude of the sum vector can be found by forming the product of the Hankel transforms of the magnitudes of the individual vectors.

Let $H_{\mathbf{v}_i}(z)$ denote the Hankel transform of the *i*th vector and $H_{\mathbf{v}_{N=k}}(z)$ denote the Hankel transform of the sum of $\mathbf{N} = k$ vectors. Then, the above description leads to

$$H_{\mathbf{v}_{N=k}}(z) = \prod_{i=1}^{k} H_{\mathbf{v}_i}(z) \tag{10}$$

Reference 9 shows that the Hankel transform of the *i*th vector can be defined as the expected value of the $rvJ_0(z\mathbf{v}_i)$ that is,

$$H_{\mathbf{v}_i}(z) = E[J_0(z\mathbf{v}_i)] \tag{11}$$

or equivalently

$$H_{\mathbf{v}_i}(z) = E[J_0(z\mathbf{a}_i\beta(t-\mathbf{t}_i))]$$
(12)

Because the individual vectors are identically distributed and the number of vectors **N** is a Poisson distributed rv, it can be shown that $H_{\mathbf{v}}(z)$, the Hankel transform of the envelope has a convenient representation in terms of the envelope of the receiver impulse response $\beta(t)$, the rate of arrival of the impulses μ , and the pdf of \mathbf{a}_i given by $f_{\mathbf{a}}(a_i)$. The actual form derived in Ref. 9 is

$$H_{\mathbf{v}}(z) = \exp\left\{-\mu \int_{0}^{\infty} f_{\mathbf{a}}(a_{i}) \int_{0}^{T} [1 - J_{0}(za_{i}\beta(s))] \, ds \, da_{i}\right\}$$
(13)

Because $\mathbf{a}_i = g(\mathbf{r}_i)$, an alternative form of Eq. (13) can be obtained in terms of the pdf of the source to receiver distance $f_{\mathbf{r}}(r_i)$ that is,

$$H_{\mathbf{v}}(z) = \exp\left\{-\mu \int_{0}^{\infty} f_{\mathbf{r}}(r_{i}) \int_{0}^{T} [1 - J_{0}(zg(r_{i})\beta(s))] \, ds \, dr_{i}\right\}$$
(14)

$$p_{\mathbf{v}}(v) = v \int_0^\infty z H_{\mathbf{v}}(z) J_0(zv) \, dz \tag{15}$$

for $v \ge 0$ and 0 otherwise.

Reference 9 also shows that the cumulative distribution function (cdf) of the envelope can be expressed as

$$F_{\mathbf{v}}(v) = v \int_0^\infty H_{\mathbf{v}}(z) J_1(zv) \, dz \tag{16}$$

for $v \ge 0$ and 0 otherwise.

The APD is defined as the probability that the envelope **v** exceeds a value v_0 and is represented as $P[\mathbf{v} > v_0]$. It is given in terms of the cdf

$$P[\mathbf{v} > v_0] = 1 - F_{\mathbf{v}}(v_0) \tag{17}$$

or equivalently for $v_0 > 0$

$$P[\mathbf{v} > v_0] = 1 - v_0 \int_0^\infty H_{\mathbf{v}}(z) J_1(zv_0) \, dz \tag{18}$$

for $v_0 > 0$.

Example. In the case of atmospheric noise we now assume that lightning discharge occur in a uniform spatial distribution about the receiver out to a specified maximum range r_m resulting in a pdf given by

$$f_{\mathbf{r}}(r_i) = 1/r_m \tag{19}$$

where $0 < r_i < r_m$.

It is further assumed that the individual impulses arrive at the receiver in accordance with an inverse propagation law $\mathbf{a}_i = k_0/\mathbf{r}_i$, where k_0 is a constant. Then, it can be shown using Eqs. (14) and (18) that the APD for v > 0 can be expressed as [9]

$$P[\mathbf{v} > v_0] = K_u / [K_u^2 + v_0^2]^{1/2}$$
(20)

where

$$K_u = \mu k_0 / r_m \int_0^T \beta(s) \, ds \tag{21}$$

This APD form has been shown to be a reasonable fit to measured atmospheric noise APDs [9,13]. Note that other assumptions on spatial distributions and propagation laws will produce other APD forms.

4. ATMOSPHERIC NOISE CHANNEL MODELS

One of the most extensively investigated impulsive noise channels is the atmospheric radio noise channel. Communications systems with center frequencies below 100 MHz must operate in the presence of atmospheric noise that arise from lightning discharges as a result of storms occurring throughout the world. Experimental data on some statistical properties of atmospheric noise can be found in several publications [5,14,15].

Receivers operating in atmospheric noise may experience two types of noise behavior, that is, 1) highly impulsive noise from local storms associated with distances that are within 1000 Km and 2) continuous noise from distant storms. With local storms radiated energy travels along the ground with low attenuation so that the receiver is subjected to strong electrical fields over a short duration. Storms that occur at ranges greater than 1000 Km propagate in modes within the earth-ionosphere cavity and arrive at the receiver with only a small portion of their original energy. In this latter case large numbers of lightning discharges occur simultaneously at various locations throughout the world causing bursts from distant storms to be numerous and overlap in time thereby producing the continuous background noise. As a result, APDs tend to have two dominant regions where the lowamplitude region is Rayleigh and arises via the central limit theorem from many independent, weak components whereas the high-amplitude region is dominated by strong distinguishable impulses that follow another distribution such as a power Rayleigh, log-normal, and so on.

The receiver bandwidth and operating frequency are also significant in characterizing impulsive noise. Very low frequency (VLF) (3-30 KHz) receivers experience significant interference as a result of the large radiated energy from the short (100 $\mu sec)$ main strokes (also referred to as the return stroke) of lightning discharges which tend to be centered around 10 KHz [45]. Inan [16] provides recent progress and results in this band. The radiated spectrum from a lightning discharge decays with frequency with the result that the predischarge consisting of several stepped discrete leaders prior to the stronger main stroke produces interference in higher frequency bands. With regard to the receiver bandwidth a receiver that uses a wide bandwidth will be subjected to individual pulses from lightning discharges that are more easily distinguished so that the noise appears impulsive. Narrow bandwidths produce overlapping pulses resulting in noise that appears more continuous.

Another mechanism that strongly impacts the time statistics of atmospheric noise is due to multiple stroking. Multiple stroking from "long" (200 μ sec) discharges usually consists of three or four return strokes spaced at about 40 msec apart. This phenomenon accounts for the departure from a random distribution in time and introduces the dependencies evident in the pulse spacing distributions.

Researchers have expended considerably less effort in modeling higher-order statistics that are important in characterizing receiver performance. One study that has investigated the effects of time dependencies in the noise pulses, involves noise measurements in the medium frequency band (300KHz-3MHz) [11]. In this case BER performance for linear and nonlinear receivers was obtained and compared by simulation using channels that included either a truncated Hall model or measured noise having pulse dependencies. The results show that when the bursty nature of the noise is neglected by assuming independent noises samples, the performance of nonlinear receivers over linear receivers is significantly greater than if pulse dependencies are incorporated in

2406 STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE

the analysis. Typically, in measured noise it is noted that nonlinearities can improve the performance of a linear receiver by an amount that is on the order of the V_d value.

With the above limited explanation of the underlying physical mechanisms, we now return to more detailed descriptions of atmospheric noise models. The models introduced in Section 2 will be extensively described and subsequently used to estimate bit error rate performance in selected cases. Other performance results on coherent and noncoherent signaling in impulsive noise can be found in Refs. 1,4,6,30,40,41.

4.1. Hall Model

The Hall model presented above is repeated here and used to compute the envelope, that is

$$\mathbf{n}(t) = \mathbf{a}(t)\mathbf{n}_g(t) \tag{22}$$

The envelope $\mathbf{v}(t)$ can now be expressed as

$$\mathbf{v}(t) = |\mathbf{a}(t)\mathbf{n}_{g}(t)| \tag{23}$$

By selecting a "two-sided" Chi distribution for $\mathbf{b}(t) = 1/\mathbf{a}(t)$, the envelope distribution fits empirical data for large values of the envelope. The pdf of $\mathbf{b}(t)$ with parameters *m* and σ^2 is then given by

$$p(b) = \frac{(\frac{m}{2})^{m/2}}{\sigma^m \Gamma(\frac{m}{2})} |b|^{m-1} \exp\left(-\frac{m}{2\sigma^2} b^2\right)$$
(24)

Note that for m = 1 the pdf of $\mathbf{b}(t)$ is Gaussian so that the pdf of $\mathbf{n}_0(t)$ is the ratio of two Gaussian processes. Hall then shows that the pdf of $\mathbf{n}_0(t)$ is given by

$$p(n_0) = \frac{\Gamma(\frac{\theta}{2})}{\Gamma(\frac{\theta-1}{2})} \frac{\gamma^{\theta-1}}{\sqrt{\pi}(n_0^2 + \gamma^2)^{\theta/2}}$$
(25)

where $\gamma = \sqrt{m} \frac{\sigma_1}{\sigma}$ and $\theta = m + 1 > 1$. For $\sigma_1 = \sigma$ the above equation is a Student *t* distribution with parameter *m* so that in the general case Hall named Eq. (25) as a generalized *t* distribution with parameters θ and γ . Letting $\mathbf{x}(t)$ and $\mathbf{y}(t)$ denote the inphase and quadrature components of $\mathbf{n}_0(t)$ respectively, Hall derives the joint pdf $p_{\mathbf{xy}}(x, y)$ from a transformation of the product of random variables in Eq. (2) resulting in

$$p_{\mathbf{xy}}(x,y) = \frac{(\theta-1)\gamma^{\theta-1}}{2\pi} \frac{1}{[x^2+y^2+\gamma^2]^{(\theta+1)/2}}$$
(26)

If we let $\mathbf{x} = \mathbf{v} \cos \psi_s$ and $\mathbf{y} = \mathbf{v} \sin \psi_s$ where \mathbf{v} and ψ_s denote the envelope and phase respectively of the noise, we can write the joint pdf of the envelope and phase as

$$p_{\mathbf{v}\psi_s}(v,\psi_s) = \frac{v(\theta-1)\gamma^{\theta-1}}{2\pi [v^2+\gamma^2]^{(\theta+1)/2}}$$
(27)

From this equation, it is seen that ψ_s is independent of **v** and has a uniform pdf in the interval $(0, 2\pi)$ so that

$$p(v) = \frac{1}{2\pi} \int_0^{2\pi} p_{\mathbf{v}\psi_s}(v, \psi_s) d\psi_s$$

= $\frac{(\theta - 1)\gamma^{\theta - 1}v}{[v^2 + \gamma^2]^{(\theta + 1)/2}}, \quad 0 \le v \le \infty$ (28)

Asymptotic forms of the pdf are

$$p(v) \approx \begin{cases} (\theta - 1)\gamma^{\theta - 1}/v^{\theta}, & \text{for } v \text{ large} \\ (\theta - 1)v/\gamma^2, & \text{for } v \text{ small} \end{cases}$$
(29)

The form of the pdf for large v is consistent with empirical data and for small v follows a limiting form of a Rayleigh distributed envelope.

The APD or exceedance distribution can be shown to be

$$P(\mathbf{v} > v_0) = \int_{v_0}^{\infty} p(v) \, dv = \frac{\gamma^{\theta - 1}}{(v_0^2 + \gamma^2)^{(\theta - 1)/2}} \tag{30}$$

To fit measured data θ is typically taken to be an integer between 2 and 5 and γ is related to the average or rms value of v for the specified value of θ . Note that if $\theta = 2$ the above equation takes the same form as Eq. (20).

The APD $P(\mathbf{v} > v_0)$ is plotted as a function of v_0/γ for $\theta = 2, 3, 4$, and 5. The Rayleigh exceedance distribution is

$$P(\mathbf{v} > v_0) = \exp\left(-\frac{v_0^2}{2\gamma^2}\right)$$

and is shown in Fig. 2 for reference purposes.³

Examination of Fig. 2 shows that longer tails are exhibited with $\theta = 2$ rather than with $\theta = 3, 4$, or 5. This behavior is consistent with impulsive noise that has a large dynamic range. A measure of the impulsiveness of the noise is V_d , the ratio of the rms to average envelope value defined as

$$V_d = 20 \log\left(\frac{\sqrt{\mu_2}}{\mu_1}\right) \tag{31}$$

where

$$\mu_j = \int_0^\infty v^j p(v) \, dv \quad j = 1, 2 \tag{32}$$



Figure 2. APD for Hall model and Rayleigh random variables.

³ These results are obtained with MATLAB.

For $\theta = 4$ and 5 $V_d = 3$ dB and 2 dB respectively and the noise is considered to be moderately impulsive. For $\theta = 3$ the second moment does not exist and for $\theta = 2$ neither the first or second moments exist. One explanation of the infinite moments can be attributed to the propagation model where the received field strength follows an inverse distance relation. Near the origin the field strength is arbitrarily large which does not correspond to a physical condition. The actual noise moments must be finite and can be forced to this condition by truncating and normalizing the envelope distribution. Thus, a Hall model envelope pdf can be developed for the truncated case and is

$$p_E(v) = \begin{cases} \frac{c(\theta - 1)\gamma^{\theta - 1}v}{[v^2 + \gamma^2]^{(\theta + 1)/2)}} & 0 \le v \le v_m \\ 0, & v > v_m \end{cases}$$
(33)

where v_m is the maximum allowed envelope level and c is selected to ensure that

$$\int_0^\infty p_E(v) \, dv = 1 \tag{34}$$

Applying the above equation, we can show that

$$c = \frac{D^{\theta - 1}}{D^{\theta - 1} - 1}$$
(35)

where $D = \sqrt{1 + (v_m/\gamma)^2}$. For $\theta = 2$ the envelope pdf of the truncated Hall model is

$$p_E(v) = \begin{cases} \frac{D\gamma}{D-1} \frac{v}{[v^2 + \gamma^2]^{3/2}}, & 0 \le v \le \gamma \sqrt{D^2 - 1} \\ 0, & v > \gamma \sqrt{D^2 - 1} \end{cases}$$
(36)

and

$$V_d = 20 \log \left(\frac{(D-1)^{3/2}}{-\sqrt{D^2 - 1} + D \ln(D + \sqrt{D^2 - 1})} \right)$$
(37)

The truncated Hall model APD is given by

$$P(\mathbf{v} > v_0) = \frac{D}{D-1} \left[\frac{1}{\left(1 + \frac{v_0^2}{\gamma^2}\right)^{1/2}} - \frac{1}{D} \right]$$
(38)

With this formulation more impulsive noise conditions can be realized and larger values of V_d can be obtained.

An advantage of the truncated Hall model is that computations can be performed for a specified V_d . For example, a highly impulsive case with $V_d = 10$ dB can be obtained by use of a normalized truncation point $v_m/\gamma = 290$. Other values for the normalized truncation point can be obtained from the plot given in Fig. 3.

4.2. Mixture Model

As described previously, mixture models are obtained by a "mixing" two simpler noise models. (See, for example, [35], [31], or [29]), that is

$$\mathbf{n}_0(t) = (1 - \varepsilon_r)\mathbf{n}_g(t) + \varepsilon_r \mathbf{n}_i(t)$$
(39)



Figure 3. V_d as a function of truncation level for Hall model.

An alternate formulation of the mixture model can be obtained in terms of the noise envelope

$$\mathbf{v}(t) = \left(\frac{\mathbf{v}_i(t) + \mathbf{v}_R(t)}{2}\right) + \left(\frac{\mathbf{v}_i(t) - \mathbf{v}_R(t)}{2}\right)\mathbf{u}(t)$$
(40)

where $\mathbf{v}_R(t)$ is a Rayleigh envelope corresponding to the Gaussian noise component, $\mathbf{v}_i(t)$ is the envelope of the impulsive noise component, and $\mathbf{u}(t)$ is +1 with probability ε and -1 with probability $1 - \varepsilon$. Thus, when $\mathbf{u}(t) = 1$, the envelope is Rayleigh distributed and given by

$$p(v_R) = \begin{cases} \frac{v_R}{\sigma^2} \exp\left(-\frac{v_R^2}{2\sigma^2}\right), & v_R \ge 0\\ 0, & \text{otherwise} \end{cases}$$
(41)

and when $\mathbf{u}(t) = -1$, the envelope is distributed in accordance with an assumed distribution. One specific impulsive envelope pdf which has been used is referred to as a generalized Laplacian [35] and is given by

$$p(v_i) = \frac{v_i^v}{2^{v-1} (P_r \sigma)^{v+1} \Gamma(v)} K_{1-v} \left(\frac{v_i}{P_r \sigma}\right), \quad v_i \ge 0$$
(42)

and zero when $v_i < 0$ where $K_{1-v}()$ is a modified Bessel function of order 1-v with v > 0 and P_r controls the ratio of the power between the impulsive and Rayleigh portions of the noise. The pdfs of the inphase and quadrature components can be determined from the joint pdf of the envelope v and phase ψ_s similar to that of Eq. (27) resulting in

$$p_{xy}(x,y) = \frac{(x^2 + y^2)^{(v-1)/2}}{2\pi (P_r \sigma)^{v+1} \Gamma(v) 2^{v-1}} K_{1-v} \left(\frac{\sqrt{x^2 + y^2}}{P_r \sigma}\right)$$
(43)

The pdf of the inphase (or quadrature) components is then, from item 6.596-3 in [12],

$$p(x) = \frac{|x|^{\nu-1/2}}{\sqrt{\pi} (P_r \sigma)^{\nu+1/2} \Gamma(\nu) 2^{\nu-1/2}} K_{1/2-\nu} \left(\frac{|x|}{P_r \sigma}\right)$$
(44)

This equation becomes a Laplace pdf when v = 1 and $P_r = 1/\sqrt{2}$ by using the relationships on page 444 in Ref. 2, that is,

$$K_{1/2}(z) = \left(\frac{\pi}{2z}\right)^{1/2} e^{-z}$$

and

$$K_{1/2-v}(z) = K_{v-1/2}(z)$$

The Laplace pdf then becomes, from Ref. 18

$$p(x) = \frac{1}{\sqrt{2\sigma^2}} \exp\left(-|x|\sqrt{\frac{2}{\sigma^2}}\right), -\infty < x < \infty$$
 (45)

Note that the Laplace pdf has zero mean and variance σ^2 so that the Gaussian and Laplacian random variables have the same mean and variance.

By computing the first and second moments of the envelope of the mixture model, the V_d ratio for the mixture model can then be determined, that is,

$$V_{d} = 20 \log \left[\frac{\sqrt{4vP_{r}^{2}\varepsilon + 2(1-\varepsilon)}}{\frac{\varepsilon P_{r}\sqrt{\pi}\Gamma(v+1/2)}{\Gamma(v)} + (1-\varepsilon)\sqrt{\frac{\pi}{2}}} \right]$$
(46)

For v = 1 this reduces to

$$V_d \mid_{v=1} = 20 \log \left[\frac{\sqrt{4P_r^2 \varepsilon + 2(1-\varepsilon)}}{\varepsilon P_r \frac{\pi}{2} + (1-\varepsilon)\sqrt{\frac{\pi}{2}}} \right]$$
(47)

With $P_r = 1$ and $\varepsilon = 0$, corresponding to the Gaussian only case, the V_d ratio is 1.05 dB. For $\varepsilon = 1$ and v = 1 the noise is impulsive according to a Laplacian distribution and $V_d = 20 \log 4/\pi = 2.1$ dB so that the noise is only mildly impulsive. More impulsive noise cases can be obtained by allowing v < 1 to be small in Eq. (46) Plots of V_d for several values of P_r and v are computed and shown in Figs. 4, 5, and 6 as a function of ε .

The APD for the mixture model can be obtained from

$$P\{\mathbf{v} > v_0\} = P\{\mathbf{v} > v_0 \mid \mathbf{u} = 1\}P\{\mathbf{u} = 1\}$$
$$+ P\{\mathbf{v} > v_0 \mid \mathbf{u} = -1\}P\{\mathbf{u} = -1\}$$
(48)

Using item 6.561–12 in Ref. 12, this can be shown to be

$$P\{\mathbf{v} > v_0\} = P\{\mathbf{v}_i > v_0\}\varepsilon + P\{\mathbf{v}_R > v_0\}(1-\varepsilon)$$

$$= \varepsilon \int_{v_0}^{\infty} \frac{v_i^v}{2^{v-1}(P_r\sigma)^{v+1}\Gamma(v)} K_{1-v}\left(\frac{v_i}{P_r\sigma}\right) dv_i$$

$$+ (1-\varepsilon) \int_{v_0}^{\infty} \frac{v_R}{\sigma^2} \exp\left(-\frac{v_R^2}{2\sigma^2}\right) dv_R = \varepsilon \left(\frac{v_0}{P_r\sigma}\right)^v$$

$$\times \frac{K_v\left(\frac{v_0}{P_r\sigma}\right)}{2^{v-1}\Gamma(v)} + (1-\varepsilon) \exp\left(-\frac{v_0^2}{2\sigma^2}\right)$$
(49)



Figure 4. V_d as a function of v and ε with $P_r = 1$.







Figure 6. V_d as a function of v and ε with $P_r = 25$.

4.3. Middleton Class A and B Models

Middleton's models are statistical physical models like that given in Eq. (1) but with more general assumptions regarding noise source spatial distributions and propagation conditions. See Refs. 25-27,38. Thus, these models are regarded as canonical because noise source spatial distributions and propagation formulas need not be explicitly specified. Both Class A and B models assume that the noise sources are Poisson distributed in space with received waveforms produced by interfering sources that are independent and Poisson distributed in time. The broadband Class B models are useful in representing atmospheric noise statistics.

For Class A noise the instantaneous amplitude has a pdf described in terms of a parameter A referred to as the impulsive index and a parameter $P_r = \sigma_g^2/\sigma_i^2$ representing the ratio of the Gaussian noise power σ_g^2 to the impulsive noise power σ_i^2 . The Class A instantaneous amplitude pdf is given by Ref. 49 as

$$p(w) = e^{-A} \sum_{m=0}^{\infty} \frac{A^m e^{-w^2/2\sigma^2 \sigma_m^2}}{m! \sqrt{2\pi \sigma_m^2 \sigma^2}}$$
(50)

where $\sigma^2 = \sigma_g^2 + \sigma_i^2$ and

$$\sigma_m^2 = \frac{\frac{m}{A} + P_r}{1 + P_r} \tag{51}$$

Middleton uses normalized instantaneous amplitudes defined as $x = w/\sigma$ so that

$$p(x) = e^{-A} \sum_{m=0}^{\infty} \frac{A^m}{m! \sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{x^2}{2\sigma_m^2}\right)$$
(52)

Small values of the parameter A produce large impulsive tails whereas large values of $A > \approx 10$ yield the limiting case of Gaussian interference. Thus, the reciprocal of A behaves as V_d in the truncated Hall and mixture models. Middleton also derives the pdf of the normalized envelope defined as $v_N = v/\sqrt{2\sigma^2}$ resulting in

$$p(v_N) = 2e^{-A} \sum_{m=0}^{\infty} \frac{A^m}{m! \sigma_m^2} v_N \exp\left(-\frac{v_N^2}{\sigma_m^2}\right), \quad v_N \ge 0$$
 (53)

and zero for $v_N < 0$.

A special case of a mixture model can be obtained by splitting the above expression into two terms as

$$p(v_N) = \frac{2e^{-A}}{\sigma_0^2} v_N e^{-v_N^2/\sigma_0^2} + 2e^{-A} \sum_{m=1}^{\infty} \frac{A^m}{m! \sigma_m^2} v_N \\ \times \exp\left(-\frac{v_N^2}{\sigma_m^2}\right), \quad v_N \ge 0$$
(54)

where

$$\sigma_0^2 = \frac{P_r}{1 + P_r} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_i^2}$$
(55)

The first term in Eq. (54) can be seen to be Rayleigh distributed whereas the second term represents an

impulsive distribution. The normalized form of the APD is obtained from

$$P\{\mathbf{v}_{N} > v_{N_{0}}\} = \int_{v_{N_{0}}}^{\infty} p(v_{N}) \, dv_{N}$$
(56)

where $v_{N_0} = N_0/\sqrt{2\sigma^2}$ resulting in

$$P\{\mathbf{v}_N > v_{N_0}\} = e^{-A} \sum_{m=0}^{\infty} \frac{A^m}{m!} \exp\left(-\frac{v_{N_0}^2}{\sigma_m^2}\right)$$
(57)

For Class B noise the instantaneous noise amplitude can be written in normalized form as

$$p(x) = \frac{e^{-x^2/W}}{\pi\sqrt{W}} \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} A^m_{\alpha} \Gamma\left(\frac{m+1}{2}\right) {}_1F_1\left(-\frac{m\alpha}{2}; \frac{1}{2}; \frac{x^2}{W}\right)$$
(58)

where ${}_{1}F_{1}$ is the confluent hypergeometric function, A_{α} is a parameter that includes the impulsive index A and other parameters that depend on the physical mechanism, α is a constant between 0 and 2 related to the noise source density and propagation law, and W is a parameter that normalizes the noise process to the energy contained in the Gaussian portion of the noise. As indicated in Ref. 42, the normalization cannot be associated with the total energy because the moments of Eq. (58) are not finite.

The pdf of the normalized envelope for Class B noise is

$$p(v_N) = \frac{2v_N}{W} \exp\left(-\frac{v_N^2}{W}\right) \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \times A^m_{\alpha} \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_1F_1\left(-\frac{m\alpha}{2}; 1; \frac{v_N^2}{W}\right), \ v_N \ge 0$$
(59)

and zero for $v_N < 0$. The normalized APD for Class B noise is

$$P\{\mathbf{v}_{N} > v_{N_{0}}\} = \exp\left(-\frac{v_{N_{0}}^{2}}{W}\right) \left[1 - \frac{v_{N_{0}}^{2}}{W} \sum_{m=1}^{\infty} \frac{(-1)^{m}}{m!} \times A_{\alpha}^{m} \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_{1}F_{1}\left(1 - \frac{m\alpha}{2}; 2; \frac{v_{N_{0}}^{2}}{W}\right)\right]$$
(60)

An alternate form for the APD of Class B noise is given by Ref. 26 as

$$P\{\mathbf{v}_{N} > v_{N_{0}}\} = 1 - \exp\left(-\frac{v_{N_{0}}^{2}}{W}\right) \frac{v_{N_{0}}^{2}}{W} \sum_{m=0}^{\infty} \frac{(-1)^{m}}{m!} \times A_{\alpha}^{m} \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_{1}F_{1}\left(1 - \frac{m\alpha}{2}; 2; \frac{v_{N_{0}}^{2}}{W}\right)$$
(61)

To see that Eqs. (60) and (61) are equivalent, we can rewrite the last equation as

$$P\{\mathbf{v}_{N} > v_{N_{0}}\} = 1 - \exp\left(-\frac{v_{N_{0}}^{2}}{W}\right) {}_{1}F_{1}\left(1;2;\frac{v_{N_{0}}^{2}}{W}\right) - e^{-v_{N_{0}}^{2}/W}\frac{v_{N_{0}}^{2}}{W}$$
$$\times \sum_{m=1}^{\infty} \frac{(-1)^{m}}{m!} A_{\alpha}^{m} \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_{1}F_{1}\left(1 - \frac{m\alpha}{2};2;\frac{v_{N_{0}}^{2}}{W}\right)$$
(62)

2410 STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE

However, from Ref. 2,

$$_{1}F_{1}(1;2;z) = \frac{e^{z}-1}{z}$$
 (63)

which allows the above equation to be written as

$$P\{\mathbf{v}_{N} > v_{N_{0}}\} = \exp\left(-\frac{v_{N_{0}}^{2}}{W}\right) - \exp\left(-\frac{v_{N_{0}}^{2}}{W}\right)\frac{v_{N_{0}}^{2}}{W}$$
$$\times \sum_{m=1}^{\infty} \frac{(-1)^{m}}{m!} A_{\alpha}^{m} \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_{1}F_{1}\left(1 - \frac{m\alpha}{2}; 2; \frac{v_{N_{0}}^{2}}{W}\right) (64)$$

For large values of the argument $z = v_{N_0}^2/W$ an approximate form of the APD, which is useful for numerical evaluation, can be obtained by replacing the confluent hypergeometric function with the expression, using Ref. 2,

$${}_{1}F_{1}(a;b;z) = \frac{\Gamma(b)}{\Gamma(a)} z^{a-b} e^{z}, \quad \text{large } z$$
(65)

Using Eq. (65) in Eq. (61) leads to

$$P\{\mathbf{v}_{N} > v_{N_{0}}\} = 1 - ze^{-z} \sum_{m=0}^{\infty} \frac{(-1)^{m} A_{\alpha}^{m}}{m!}$$
$$\times \Gamma\left(1 + \frac{m\alpha}{2}\right) \frac{\Gamma(2)}{\Gamma(1 - \frac{m\alpha}{2})} z^{-\frac{m\alpha}{2} - 1} e^{z}$$
$$= -\sum_{m=1}^{\infty} (-1)^{m} \frac{A_{\alpha}^{m}}{m!} \frac{\Gamma(1 + \frac{m\alpha}{2})}{\Gamma(1 - \frac{m\alpha}{2})} z^{-m\alpha/2}, \quad \text{large } z \quad (66)$$

From Ref. 2, we can use the identities

$$\Gamma\left(1+\frac{m\alpha}{2}\right) = \frac{m\alpha}{2}\Gamma\left(\frac{m\alpha}{2}\right) \tag{67}$$

and

$$\Gamma\left(\frac{m\alpha}{2}\right)\Gamma\left(1-\frac{m\alpha}{2}\right) = \frac{\pi}{\sin\pi\alpha\frac{m}{2}}$$
(68)

in Eq. (66) resulting in

$$P\{\mathbf{v}_N > v_{N_0}\} = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{A_{\alpha}^m}{(m-1)!} \frac{\alpha}{2\pi}$$
$$\times \Gamma^2\left(\frac{m\alpha}{2}\right) \sin\left(\pi\alpha \frac{m}{2}\right) z^{-m\alpha/2}, \quad \text{large } z \qquad (69)$$

An example of the APD of Middleton's Class B noise for $V_d = 10 \text{ dB}$ is given below in Section 4.5.

4.4. Empirical First-Order Model

The empirical first-order (simulation) noise model is based upon the Crichlow graphical model for the APD of the noise envelope [7]. This model uses a special type of probability graph paper; one on which the power Rayleigh functions plot as straight lines.⁴ The coordinate transformations for the probability paper is found by considering the APD of a normalized Rayleigh function

$$\frac{n}{\sqrt{\mu_2}} = \operatorname{APD}(w) = e^{-w^2}$$
(70)

We use the transformation

$$x' = -20\log(-\ln w)$$
(71)

$$y' = 20 \log \frac{n}{\sqrt{\mu_2}} \tag{72}$$

On this probability paper, the APD of atmospheric noise can be represented by a three-section curve as shown in Fig. 7. The lower region of the curve, representing random low-amplitude envelopes of high probability, approaches a Rayleigh distribution. Hence, it can be approximated by a straight line (R). The higher region of the curve, representing high-impulsive envelopes of low probability, approaches a power Rayleigh distribution. It can also be approximated by a straight line (PR). The center region of the curve corresponds to a circular arc tangent to the two straight lines. The circular arc is also tangent to the line (T) which is parallel to line (BI), which bisects the acute angle formed by the intersection of the Rayleigh and power Rayleigh lines. Four parameters are necessary to specify a unique pair of lines and an arc. They are:

- 1. Slope of the power Rayleigh line;
- 2. Point through which the power Rayleigh line passes;
- 3. Point through which the Rayleigh line passes (the slope is known to be $\frac{-1}{2}$);
- 4. Parameter determining the radius of the circular arc.

Crichlow defined the four parameters as follows:

- 1. X = -2s, where s is the slope of the power Rayleigh line;
- 2. C(dB) = the dB difference between the power Rayleigh line and the Rayleigh line at p = 0.01;
- 3. A (dB) =the dB value of the Rayleigh line at p = 0.5;
- 4. B(dB) = the dB difference between the y' -axis intercepts of lines (BI) and (T).

Experimentally measured APDs indicate that the parameter B is linearly related to first order to the parameter X by

$$B = 1.5(X - 1) \tag{73}$$

Thus, once the values of parameters X, C, and A are known, a unique APD can be constructed. The X, C, and A parameters vary according to the V_d value of the APD. Wilson [47] has calculated the values of these parameters for V_d values of 4.0 through 30.0. Parameter values for $V_d = 2.0$ and 3.0 were determined by extrapolating upon Wilson's values along with experimental curve fitting. The above transformations were thoroughly tested with a total of 20,000 samples, for each of ten different V_d values. The APDs were then determined by experimental

⁴ A Rayleigh function plots as a straight line with slope = -.5.



Figure 7. Crichlow APD model.

measurements. This can be seen in Fig. 8 for V_d ratios of 6, 10, and 14 dB. It is clear from the figure that more than 20,000 samples should be used to test the APD curves for values of Δ greater than 20.

4.5. Comparison of Selected APD Results for Different Models

In this section the Middleton Class B model, the truncated Hall model, and the empirical model are compared with CCIR 322 data [5]. The main point of this section is to show that numerous models can be used to obtain good fits to measured APD data and that no matter what model is selected, error rate performance estimates will be the same as long as the noise samples are independent.⁵

Figure 9 shows a comparison of the APD curves for the truncated Hall model with $V_d = 10$ dB, the Middleton Class B model with $A_{\alpha} = 1$, $\alpha = 1$, and W = 0.007, and CCIR 322 data with $V_d = 10$ dB. For reference purposes, the Rayleigh APD is also shown; it is known that the Rayleigh distribution corresponds to $V_d = 1.05$ dB. The



Figure 8. Verification of Crichlow APD using Wilson's parameters.

curves are plotted on semilog paper where the vertical scale is in dB above the rms level. For the trancated Hall case, the rms level is $\gamma\sqrt{D-1}$ and for the Rayleigh case the rms level is $\sqrt{2\gamma}$. The parameter W provides the normalization for the Class B Middleton model. A comparison of the empirical model for $V_d = 10$ dB

⁵ Atmospheric noise measured data reveals that noise samples are correlated because of multiple lightning discharges. Nevertheless, it can be shown that the APD is essentially the same; higher order statistics will, however, be affected, although this behavior is rarely modeled. See Ref. 9.



Figure 9. Comparison of some atmospheric noise models.

is presented in Fig. 8. Good fits with CCIR 322 data are obtained in these particular examples, especially for envelope values in dB above rms that are below about 20 dB. Computational problems arise with both the truncated Hall model and the Class B model for large envelope values. In the trancated Hall case the curve stops when the truncation point is exceeded. In the Class B case the exact formulation in Eq. (61) converges poorly so that the approximation of Eq. (66) is involved for z > 100. The Class B parameters were selected to provide good but not necessarily an optimum fit to the CCIR 322 data.

5. DETECTOR STRUCTURES IN NON-GAUSSIAN NOISE

At this point it is apparent that numerous models exist to represent non-Gaussian noise. The development of optimal structures would, in principle, require computation of the likelihood ratio or log-likelihood ratio. Because of the non-Gaussian noise, however, each model can in general lead to a unique optimum detector structure. This section describes a few of the common cases which have been derived and built. As indicated below in Section 5.1, an optimal detector structure under the conditions of small snr leads to the general form of a nonlinearity followed by a linear correlator or matched filter demodulation for making a decision. Figure 10 shows the general structure. A wide variety of nonlinearities have been used, including hard limiters, soft clippers, hole punchers, and logarithmic devices [10].

Section 5.2 shows one detector structure that has been developed based upon Hall's model. Finally, in Section 5.3, some commonly-used ad-hoc structures are presented.



Figure 10. Common detector structure for impulsive noise.

5.1. Weak Signal Detection

A commonly derived receiver structure for the case of a weak signal is referred to as the locally optimum Bayes' detector (LOBD).(The LOBD is also called a low snr detector or a threshold receiver.) References 17, 19, and 40 give up-to-date derivations of this detector, although Ref. 3 gives one of the earliest presentations. Following the presentations in Refs. 3 and 40, we define two hypotheses as

$$\vec{\mathbf{y}} = \vec{\mathbf{u}}_i + \vec{\mathbf{n}}, \quad i = 0, 1 \tag{74}$$

where each vector contains k samples. The likelihood ratio is given by

$$L(\vec{y}) = \frac{p_1(\vec{y})}{p_0(\vec{y})} = \frac{p_n(\vec{y} - \vec{u}_1)}{p_n(\vec{y} - \vec{u}_0)}$$
(75)

We now expand the pdf $p_n(\vec{y} - \vec{u})$ in a vector Taylor series and ignore, for small signals, terms of degree two and higher. This results in

$$p_{\mathbf{n}}(\vec{y} - \vec{u}_i) \simeq p_{\mathbf{n}}(\vec{y}) - \sum_{j=1}^k \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_i} u_{ij}$$
(76)

Substituting this expression into the likelihood ratio above yields

$$L(\vec{y}) = \frac{p_{\mathbf{n}}(\vec{y}) - \sum_{j=1}^{k} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_{i}} u_{1j}}{p_{\mathbf{n}}(\vec{y}) - \sum_{i=1}^{k} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_{i}} u_{0j}}$$
(77)

Dividing the numerator and denominator by $p_{\mathbf{n}}(\vec{y})$ results in

$$L(\vec{y}) = \frac{1 - \sum_{j=1}^{k} \frac{1}{p_{\mathbf{n}}(\vec{y})} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_{i}} u_{1j}}{1 - \sum_{j=1}^{k} \frac{1}{p_{\mathbf{n}}(\vec{y})} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_{i}} u_{0j}}$$
$$= \frac{1 - \sum_{j=1}^{k} \frac{d}{dy_{j}} \ln p_{\mathbf{n}}(y_{j}) u_{1j}}{1 - \sum_{j=1}^{k} \frac{d}{dy_{j}} \ln p_{\mathbf{n}}(y_{j}) u_{0j}}$$
(78)

The unity constant does not affect the decision and can be ignored leading to the decision to choose H_1 if⁶

$$-\sum_{j=1}^{k} \frac{d}{dy_{j}} \ln p_{\mathbf{n}}(y_{j}) u_{1j} > -\sum_{j=1}^{k} \frac{d}{dy_{j}} \ln p_{\mathbf{n}}(y_{j}) u_{0j}$$
(79)

The receiver structure for this case is depicted in Fig. 11 and is often referred to as a threshold receiver, that is,

 $^{6}\,\mathrm{These}\,$ results assumes equal a priori probabilities and uniform costs.

STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE 2413



Figure 11. LOBD detector.

LOBD, because the signal is assumed to be small. It is apparent that the structure of Fig. 11 is an example of that of Fig. 10.

It is instructive to examine the form of Fig. 11 when the noise is white and Gaussian. In this case, the density is

$$p_{\mathbf{n}}(\vec{y}) = \frac{1}{(2\pi\sigma^2)^{k/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^k y_j^2\right), \quad (80)$$

$$\ln p_{\mathbf{n}}(\vec{y}) = -\frac{k}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^k y_j^2,$$
(81)

and

$$-\frac{d}{dy_i}\ln p_{\mathbf{n}}(\vec{y}) = \frac{y_j}{\sigma^2}$$
(82)

In other words, the nonlinearity, in this case, is linear. This is comforting for we know that the optimum detector in white Gaussian noise is linear.

An example of a non-Gaussian noise distribution for which the detector can be solved in its entirety is the Laplace distributed noise. In this case, k independent noise samples comprise the vector $\vec{\mathbf{n}}$ so that the received signal vector can be represented by

$$\vec{\mathbf{y}} = \vec{\mathbf{u}}_i + \vec{\mathbf{n}} \tag{83}$$

where \vec{u}_i , i = 0, 1 is the transmitted signal vector. The pdf of \vec{n} is given by

$$p(\vec{n}) = \frac{1}{\sqrt{2\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}} \sum_{j=1}^k |n_j|\right)$$
(84)

Under hypothese H_i , i = 0, 1 the pdf of \vec{y} becomes

$$p(\vec{y}_j) = \frac{1}{\sqrt{2\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}} \sum_{j=1}^k |y_j - u_{ij}|\right)$$
(85)

The log-likelihood ratio can now be computed as

$$\ell(\vec{y}) = \ln \frac{p_1(\vec{y})}{p_0(\vec{y})} = \sqrt{\frac{2}{\sigma^2}} \sum_{j=1}^{k} [|y_j - u_{0j}| - |y_j - u_{1j}|]$$
(86)

and the decision rule is to choose H_1 when⁷

$$\sum_{j=1}^{k} |y_j - u_{0j}| > \sum_{j=1}^{k} |y_j - u_{1j}|$$
(87)

Continuing with the example, if we let $u_{1j} = c$ and $u_{0j} = -c$ for all *j* and define the function $g(y_j)$ as

$$g(y_j) = |y_j + c| - |y_j - c|,$$
 (88)

then the receiver structure can be obtained and is shown in Fig. 12 where the Laplace noise nonlinearity is shown in Fig. 13. This detector structure is of the general form of Fig. 11 but in this example it is not necessary to impose the weak signal restriction.

5.2. Hall's Log-Correlator

Hall derived the optimum receiver for Hall model noise by forming a likelihood ratio computed from the pdf of the complex envelope of the noise as shown in [13]. A few definitions are required before the pdf of the complex noise envelope can be obtained. Assume samples of the received noise are taken at a spacing Δt and let $\mathbf{v}_g(t)$ be the complex envelope of the noise $\mathbf{n}_g(t)$ in Eq. (2). We further assume that $\mathbf{v}_g(t)$ has the covariance

$$E\{\mathbf{v}_{g}(t_{i})\mathbf{v}_{g}^{*}(t_{j})\} = N_{0}\delta_{ij}, \quad i, j = 1, \dots, k$$
(89)



Figure 12. Detector structure for laplace noise.



Figure 13. Laplace noise nonlinearity.

 $^7\,{\rm This}$ decision rule assumes equal a priori probabilities and uniform costs.

2414 STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE

where k represents the number of received complex samples. If the noise signal has a duration T and 2B is the RF bandwidth of the receiver front-end, $k \approx 2BT$, and a good approximation of the complex noise envelope results.

The process $\mathbf{b}(t)$, which is the reciprocal of the slowly varying modulating process $\mathbf{a}(t)$ in Eq. (2) has a zero mean with a pdf given by Eq. (24) and a covariance that is assumed to be

$$E\{\mathbf{b}(t_i)\mathbf{b}(t_j)\} = \frac{B_0}{2}\delta_{ij}, \quad i, j = 1, \cdots, k$$
(90)

Let \vec{z} be a vector of k complex envelope samples for the Hall model in Eq. (2) for the special case $\theta = 2$. Hall then shows that the pdf of the noise in the independent sample case can be expressed as

$$p(\vec{z}) = \left(\frac{\gamma_2}{2\pi}\right)^k \prod_{j=1}^k \frac{1}{[|z_j|^2 + \gamma_2^2]^{3/2}}$$
(91)

where $\gamma_2^2 = (N_0 \Delta t)/B_0$. Under hypothesis H_i , i = 0, 1, the received signal vector $\vec{\mathbf{y}}$ can be written as a sum of the complex signal vector \vec{u}_i and the noise vector $\vec{\mathbf{z}}$ as

$$\vec{\mathbf{y}} = \vec{u}_i + \vec{\mathbf{z}}, \quad i = 0, 1 \tag{92}$$

and from Eq. (91) the pdf of $\vec{\mathbf{y}}$ is given by

$$p(\vec{y}) = \left(\frac{\gamma_2}{2\pi}\right)^k \prod_{j=1}^k \frac{1}{[|y_j - u_{ij}|^2 + \gamma_2^2]^{3/2}}$$
(93)

The log-likelihood ratio can then be written as

$$\ell(\vec{y}) = \ln \frac{p_1(\vec{y})}{p_0(\vec{y})} = \sum_{j=1}^k \ln[|y_j - u_{0j}|^2 + \gamma_2^2]^{3/2} - \sum_{i=1}^k \ln[|y_j - u_{1j}|^2 + \gamma_2^2]^{3/2}$$
(94)

Note that the power 3/2 has no effect on the decision so that an equivalent rule is to choose H_1 when

$$\sum_{j=1}^{k} \ln[=|y_j - u_{0j}|^2 + \gamma_2^2] > \sum_{j=1}^{k} \ln[|y_j - u_{1j}|^2 + \gamma_2^2] \quad (95)$$

Hall's log-correlator is depicted at baseband in Fig. 14. Note that the term γ_2^2 is referred to as the bias and must in general be estimated. Although the log-correlator receiver was developed for $\theta = 2$, Hall shows that the decision rule of Eq. (95) is the same when γ_2^2 is replaced by $\gamma^2 = m\gamma_2^2$.

Exact error rate computations for this receiver are intractable even with a central limit theorem argument which requires only the first and second moments of the log of the decision variable. Instead Hall computes bounds on the probability of error. Below, in Section 6.2, we present



Figure 14. Log-correlator receiver.

error rate results for the log-correlator and other receiver structures obtained by simulation.

5.3. Ad Hoc Receiver Structures

Optimal Zero Memory Nonlinear (ZMNL) devices are often difficult to implement in practice. Thus, more practical, yet suboptimal, nonlinearities are used. Figure 15 shows the three most common ad hoc nonlinearities. In Fig. 15a the transfer characteristics of a hole puncher are shown. In this case, the input signal is passed undistorted as long as its envelope is smaller than a threshold t_h . If any samples of the envelope are larger than t_h , these samples are totally suppressed.

In Fig. 15b, the transfer characteristics for a clipper ZMNL device is shown. Its characteristics are similar to a hole puncher except that if the envelope exceeds t_h , a constant output proportional to t_h is available rather than a total suppression of the signal. By allowing t_h to get very small relative to the signal plus noise envelope, the clipper approximates a hard limiter, whose characteristics are shown in Fig. 15c.

6. SELECTED EXAMPLES OF NOISE MODELS, RECEIVER STRUCTURES, AND ERROR RATE PERFORMANCE

Sample results are presented in this section for several noise models and corresponding receiver structures. No attempt is made to be exhaustive because in general, every noise model yields a different receiver structure and its associated error rate performance. Conversely, as mentioned previously, if the APD of the noise models is essentially the same, then the resulting error rate performance for any *specific* receiver structure will be similar for all the noise models, assuming independence of the noise samples.

The results provided here illustrate the available improvement in error rate performance when a nonlinear receiver is utilized in the presence of impulsive noise. Sample cases presented below include:

• Bandpass limiter and linear correlator in Gaussian noise, outlined in Section 6.1.



Figure 15. Ad Hoc zero memory nonlinearities used on atmospheric noise channel receivers.

- Bandpass limiter, linear correlator, and logcorrelator performance in truncated Hall noise with a V_d of 10 dB, presented in Section 6.2.
- Weak signal receiver with Middleton noise, presented in Section 6.3, and
- Hole puncher and soft clipper in empirically generated noise with a V_d of 6 dB, detailed in Section 6.4.

In the impulsive noise cases multiple samples per symbol are assumed to allow the matched filter or linear correlator to average the residual noise following the nonlinearity. It can be shown that the case of a single sample per symbol provides no advantage from use a nonlinearity. Error rate performance is obtained by simulation for coherent antipodal signaling with the number of samples per symbol denoted by NSAM and the number of symbols used denoted by NSYM. For an RF receiver filter bandwidth denoted by 2B and a symbol duration of T, then NSAM = 2BT so that the signal to noise ratio SNR is related to the energy contrast ratio E_b/N_0 by SNR*NSAM = E_b/N_0 .

6.1. Bandpass Limiter and Linear Correlator in Gaussian Noise

Because the optimum receiver in independent Gaussian noise is a linear receiver, use of a bandpass limiter (or any other nonlinearity) in this case prior to the linear correlator will therefore degrade the error rate performance. Figures 16 and 17 show the BPL/linear correlator and linear correlator receivers respectively. These receivers structures were used in a simulation to estimate error rate performance. The simulated transmitted sequence was assumed to be symbols with values that are ± 1 . White Gaussian noise was added to the transmitted sequence and then input to each receiver. Using MATLAB, error rate performance shown in Fig. 18 was obtained for NSAM = 10 and NSYM = 5000. The theoretical error rate is expressed as

$$P_e = \frac{1}{2} \operatorname{erfc}\left(\sqrt{\frac{E_b/N_0}{\text{NSAM}}}\right)$$
(96)

The simulation shows good agreement between the linear correlator simulation and the theoretical performance and



Figure 16. Bandpass limiter/linear correlator receiver.



Figure 17. Linear correlator receiver.



Figure 18. Error rate performance for a linear and BPL/linear receiver in Gaussian noise.

that the BPL performance is approximately 1-dB poorer than the linear receiver for larger values of E_b/N_0 .

6.2. Bandpass Limiter, Linear Correlator, and Log-Correlator in Truncated Hall Noise

To generate the noise samples, in-phase and quadrature samples of the Hall model noise are computed from

$$\mathbf{n}_{0r} = |\mathcal{T}| \cos \mathbf{u} \tag{97}$$

and

$$\mathbf{n}_{0i} = |\mathcal{T}|\sin\mathbf{u} \tag{98}$$

where \mathcal{T} is a random variable from a generalized *t* distribution and **u** is uniformly distributed in the interval $(-\pi, \pi)$. The random variable \mathcal{T} is defined by

$$\mathcal{T} = \frac{\mathcal{G}}{\sqrt{\frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_{i}^{2}}}$$
(99)

where \mathcal{G} is a zero mean Gaussian random variable with variance σ_1^2 and the \mathbf{x}_i are independent zero mean Gaussian distributed random variables with variance σ^2 . Note that $\sum_{i=1}^{m} \mathbf{x}_i^2$ is chi-squared with m degrees of freedom and is independent of \mathcal{G} . Therefore $|\mathcal{T}|$, can be generated from the ratio of a Rayleigh distributed random variable \mathbf{r} to the square root of a chi-squared random variable with m degrees of freedom. \mathbf{r} in turn can be obtained from a transformation of a uniformly distributed random variable \mathbf{u} on the interval (0, 1) by

$$\mathbf{r} = [-2\sigma_1^2 \ln \mathbf{u}]^{1/2} \tag{100}$$

The Hall bias term γ is computed from $\gamma^2 = m\sigma_1^2/\sigma^2$. The above procedure can also be used to compute noise samples from a truncated Hall distribution. In such a case, the samples obtained from the Hall model noise with $\theta = 2$ are generated and samples whose amplitude exceed the truncation value v_m are discarded.

Before presenting the simulation results for the truncated Hall model, a result for one sample per symbol in a linear receiver is derived. For symbols **u** that assume the values of $\pm a$ with equal probability, the error rate can be expressed as

$$P_e = P\{\mathbf{u} > 0 \mid \mathbf{u} = -a + \mathbf{n}\}$$
(101)

where **n** is the additive truncated Hall model noise with parameter γ , v_m , and $\theta = 2$. Because the pdf of **n** is

$$p(n) = \frac{1}{\pi \gamma [(\frac{n}{\gamma})^2 + 1]}$$
(102)

it follows that

$$P_{e} = \int_{0}^{n_{m}} \frac{du}{\pi \gamma [(\frac{u+a}{\gamma})^{2} + 1]}$$
(103)

where n_m is the maximum value of the noise corresponding to the truncation point. For high V_d , the truncation point can be approximated as infinite so that the resulting error probability becomes

$$P_e \approx \frac{1}{2} - \frac{1}{\pi} \tan^{-1}\left(\frac{a}{\gamma}\right) \tag{104}$$

For a = 1, the average transmitted power is unity and the noise power is $\gamma^2(D-1)$ so that the parameter γ can be replaced by

$$\gamma = \sqrt{\frac{1}{(D-1)E_b/N_0}} \tag{105}$$

Figure 19 shows⁸ the error rate results for the linear correlator, BPL/linear correlator, and log-correlator



Figure 19. Error rate performance for various receivers in truncated hall noise.

⁸ Note that some scatter occurs at low estimated error probabilities as a result of an insufficient number of samples for the estimate.

receivers for NSAM = 10 and NSYM = 10000; for reference the theoretical error rate for one sample per symbol and a = 1 is also shown. The curves are obtained by simulation using MATLAB. It can be seen that the log-correlator receiver is the best and the BPL/linear correlator is slightly degraded. Conversely, both nonlinear receivers are significantly better than the linear correlator. These results neglect pulse dependencies from multiple stroking where the bursty nature of measured atmospheric noise tends to limit the performance improvement of nonlinear receivers to an amount that is on the order of the V_d value [15,11].

6.3. WEAK SIGNAL DETECTOR IN MIDDLETON NOISE

Returning to the result of Fig. 11, Spaulding and Middleton, in Refs. 37 and 40, provide a general result for the error rate performance of antipodal signals in Middleton's Class A noise as

$$P_e = \frac{1}{2} \operatorname{erfc} \left(\sqrt{k\Gamma_o f/2} \right), \Gamma_o f \ll 1$$
(106)

where Γ_o is the signal to noise ratio and *f* is defined as

$$f = \int_{-\infty}^{\infty} \left[\frac{d}{dy} \ln p_n(y) \right]^2 p_n(y) \, dy \tag{107}$$

Performance results for noncoherent signaling in Middleton's Class A noise and weak signals are provided by Spaulding and Middleton in Refs. 37 and 41, leading to an approximate error rate for the LOBD as

$$P_e = \frac{1}{2} \exp\left(-\frac{k\Gamma_o f}{4}\right), \Gamma_o f \ll 1$$
(108)

6.4. Hole Puncher and Clipper in Empirically Generated Noise

In this section the noise is generated using the empirical method described in Section 4.4 using $V_d = 6 \text{ dB}$. The error rate performance, displayed in Figs. 20 and 21, for the soft clipper and hole puncher receivers respectively, are obtained parametric in the threshold t_h in dB above the average envelope. Although these figures are displayed using MATLAB, the original data is derived from a simulation reported on in Ref. 36. The simulated transmitted signal is Minimum Shift Keying (MSK) with a two-sided bandwidth of 900 Hz and a signaling rate of 100 bps. The best threshold for the soft clipper is 0 dB with little degradation for a setting of 3 dB. The best threshold for the hole puncher is 3 dB for high signal to noise ratios. It should be noted that the performance for both nonlinearities obtained with their optimum thresholds is about the same, as evident in Fig. 22. For reference, a Gaussian error rate curve for coherent detection of MSK is displayed in all three figures.

7. CONCLUSIONS

This article has provided a brief introduction on impulsive noise physical phenomena, impulsive noise statistics,



Figure 20. Probability of error as a function of SNR per bit with clipper threshold as a parameter.

2

appropriate utilized receiver structures and some selected BER performance results. A few significant points that have been presented are now emphasized:

- 1. There is no single noise model that can be used as a representation of all impulsive noise channels.
- 2. Physical measurements of the impulsive noise characteristics are generally needed to understand the behavior of the noise allowing noise models, receiver structures, and performance estimates to be obtained.



- 3. Optimum receiver structures are nonlinear but are "optimum" only in terms of the noise model assumed.
- 4. Specific channels require knowledge of higher order statistics to completely characterize the noise process and produce reliable performance estimates.

It is anticipated that future work on the statistical characterization of impulsive noise for various communications channels will involve new measurements and continued emphasis on first order statistical models that



Figure 22. Comparison of bit error rates for two nonlinearities.

are analytically tractable. It is also expected that significant research will focus on the time statistics of the noise for the channel under investigation. Gratifying results may very well be attained by continued emphasis on the underlying physical process such as that developed for statistical physical models.

BIOGRAPHIES

Arthur A. Giordano formed AG Consulting, LLC in June 2001 to consult in the field of military and commercial communications. Specific consulting activities have included: 1) investigation of secure net broadcast for voice and data to evaluate the potential interoperability of commercial cellular systems including CDMA, GSM, TDMA, and next generation (3G) cellular systems with planned military systems; 2) investigation and documentation on a comparison of SMR, CDMA, 1xRTT, and GPRS 3) expert witness on cases involving RF coverage and next generation CDMA; and 4) development of a satellite channel simulator. From 1993 to June 2001, he worked as a wireless manager for GTE and later Verizon Laboratories and was primarily involved in the analysis, design, and deployment of cellular and fixed wireless networks. From 1985 to 1993, he was a vice president at CNR responsible for a critical government program to provide secure, formatted message traffic for a user network employing multimedia communications assets consisting of landlines, satellites, and HF radios. Prior to 1985, he held several engineering positions encompassing communications network architecture development, spread spectrum communications, modulation and coding designs, adaptive signal processing, and measurement and modeling of atmospheric noise. He has a Doctorate from the University of Pennsylvania in EE and MS and BS degrees in EE from Northeastern University. He has published numerous technical articles, holds two patents, has coauthored a book entitled Least Square Estimation with Applications to Digital Signal Processing and is currently coauthoring a text on Detection and Estimation Theory.

Thomas A. Schonhoff received his Bachelor's degree from M.I.T., his Master's degree from Johns Hopkins University, Baltimore, Maryland, and his Ph.D. from Northeastern University, Boston, Massachusetts. He has worked at six different corporations, although he has been with LinCom Corporation (now Titan System Corporation, Communication and Software Solutions Division) since 1985. For the last 21 years, Dr. Schonhoff has also taught graduate courses as an adjunct at Worcester Polytechnic Institute, Massachusetts.

BIBLIOGRAPHY

- 1. B. Aazhang and H. V. Poor, performance of DS/SSMA communications in impulsive channels. II hard-limiting correlation receivers, *IEEE Trans. Comm.* **36**(1): 88 an 1988.
- M. Abramowitz and I. A. Stegun, ed., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, U.S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series 55, 1968.
- O. Y. Antonov, Optimum Detection of Signals in Non-Gaussian Noise, *Radio Eng. Elec. Physics* 12: 541–548 April 1967.
- S. Buzzi, E. Conte, and M. Lops, Optimum Detection over Rayleigh Fading Dispersive Channels, with Non-Gaussian Noise, *IEEE Trans. Comm.* 45(9): 1061–1069 (Sept. 1997).
- 5. CCIR Report 322, World distribution and characteristics of atmospheric radio noise, International Radio Consultative Committee, Genevsa, Feb. 7, 1964.

- E. Conte, M. DiBisceglie, and M. Lops, Optimum detection of fading signals in impulsive noise, *IEEE Trans. Comm.* 43(2): Part 3 869–876 (Feb.-March-April 1995).
- W. Q. Crichlow et al., Determination of the amplitudeprobability distribution of atmospheric radio noise from statistical moments, J. Res. Natl. Bureau Stand. — D: Radio Propag. 64D: 49-56 (1960).
- K. Furutsu and T. Ishida, On the theory of amplitude distribution of impulsive noise, J. Appl. Phys. 32: 1206–1221 (July 1961).
- A. Giordano and F. Haber, Modeling of atmospheric noise, Radio Sci. 7(11): 1011-1023 (1972).
- A. A. Giordano and H. E. Nichols, Simulated error rate performance of nonlinear receivers in atmospheric noise, *National Telecommunications Conference*, 1977.
- A. A. Giordano et al., Measurement and statistical analysis of wide-band MF atmospheric radio noise, 2, impact of data on bandwidth and system performance, *Radio Sci.* 21(2): 203-222 (Mar.-Apr. 1986).
- I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series,* and *Products*, corrected and enlarged edition, prepared by A. Jeffrey, Academic Press, 1980.
- H. M. Hall, A New Model for 'Impulsive' Phenomena: Application to Atmospheric-Noise Communication Channels, Ph.D. dissertation, Stanford University, CA, Aug. 1966.
- 14. J. R. Herman et al., Measurement and statistical analysis of wide-band MF atmospheric radio noise, 1 structure and distribution and time variation of noise power, *Radio Sci.* 21(1): 25-46 (Jan.-Feb. 1986).
- J. R. Herman et al., considerations of atmospheric noise effects on wideband MF communications, *IEEE Commun.* Mag. 24-29 (Nov. 1983).
- U Inan, Holographic Array for Ionospheric Lightning (HAIL), http://www-star.stanford.edu/vlf/
- S. A. Kassam, Signal Detection in Non-Gaussian Noise, Springer-Verlag, New York, 1988.
- S. M. Kay Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory, Prentice-Hall, Englewood Cliffs, NJ, 1998.
- K. J. Kerpez and K. Sistanizadeh, High bit rate asymmetric digital communications over telephone loops, *IEEE Trans. Comm.* 43(6): 2038–2049 (June 1995).
- K. Kumozaki, Error correction performance in digital subscriber loop transmission systems, *IEEE Trans. Comm.* 39(8): 1170-1174 (Aug. 1991).
- W. R. Lauber and J. M. Bertrand, Statistics of motor vehicle ignition noise at VHF/UHF, *IEEE Trans. Electromag. Compat.* 41(3): 257-259 (Aug. 1999).
- R. N. McDonough and A. D. Whalen, Detection of Signals in Noise, 2nd ed., Academic Press, New York, 1995.
- 24. S. McLaughlin, The Non-Gaussian Nature of Impulsive Noise in Digital Subscriber Lines, http://www.ee.ed.ac.uk/bin/ pubsearch?McLaughlin, Jan. 2000.
- D. Middleton, Statistical-physical models of electromagnetic interference, *IEEE Trans. Electromagn. Compat.* EMC-19: 106-127 (Aug. 1977).
- 26. D. Middleton, Non-Gaussian noise models in signal processing for telecommunications: New methods and results for

class a and b noise models, *IEEE Trans. Inform. Theory* **45**(4): 1129–1149 (May 1999).

- D. Middleton and A. D. Spaulding, Elements of Weak Signal Detection in Non-Gaussian Noise Environments, Advances in Statistical Signal Processing, Vol. 2, Signal Detection, H. V. Poor and J. B. Thomas editors, JAI Press, Inc., Greenwich, CT, 1993.
- D. Middleton, Procedures for determining the parameters of the first-order canonical models of class a and class b electromagnetic interference, *IEEE Trans. Electromagn. Compat.* EMC-21(3): 190-208 (Aug. 1979).
- J. H. Miller and J. B. Thomas, Detection of signals in impulsive noise modeled as a mixture Process, *IEEE Trans. Comm.* Com-24: 559-563 (May 1976).
- S. Miyamoto, M Katayama, and N Morinaga, Performance analysis of QAM systems under class a impulsive noise environment, *IEEE Trans. Electromagn. Compat.* 37(2): 260-267 (May 1995).
- J. W. Modestino, Locally optimum receiver structures for known signals in non-Gaussian narrowband Noise, 13th Annual Allerton Conference on Circuit and System Theory, Oct. 1975.
- J. B. O'Neal Jr., Substation noise at distribution-line communications frequencies, *IEEE Trans. Electromagn. Compat.* 30(1): 71-77 (Feb. 1988).
- 33. John G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
- D. H. Sargrad and J. W. Modestino, Errors-and-erasure coding to combat impulse noise on digital subscriber loops. *IEEE Trans. Comm.* 38(8): 1145–1155 (Aug. 1990).
- 35. T. A. Schonhoff, A. A. Giordano, and Z. McC, Huntoon, Analytical representation of atmospheric noise distributions constrained in V_d , 1977 International Conference on Communications.
- T. A. Schonhoff, A differential GPS receiver system using atmospheric noise mitigation techniques, *Proceedings of the ION GPS*-91, September 1991.
- A. D. Spaulding and D. Middleton, *Optimum Reception in an Impulsive Interference Environment*, U.S. Dept. of Commerce OT Report 75–67, June 1975.
- A. D. Spaulding, Stochastic modeling of the electromagnetic interference environment, 1977 International Conference on Communications, pp. 42.2–114, 42.2–123.
- A. D. Spaulding, C. J. Roubique, and W. Q. Crichlow, Conversion of the amplitude probability distribution function for atmospheric radio noise from one bandwidth to another, J. Research NBS 66D(6): 713-720 (Nov.-Dec.-Feb. 1962).
- A. D. Spaulding and D. Middleton, Optimum reception in an impulsive interference environment—part I: Coherent detection, *IEEE Trans. Commun.* COM-25: 910-923 (Sept. 1977).
- A. D. Spaulding and D. Middleton, Optimum reception in an impulsive interference environment — part II: Incoherent detection, *IEEE Trans. Commun.* COM-25: 924–934 (Sept. 1977).
- 42. A. D. Spaulding, Optimum threshold signal detection in broad-band impulsive noise employing both time and spatial sampling, *IEEE Trans. Commun.* COM-29: (Feb. 1981).
- 43. URSI, Review of Radio Science 1981-1983.
- 44. URSI Commission E.

2420 STATISTICAL MULTIPLEXING

- 45. A. D. Watt, VLF Radio Engineering, Pergamon, London, 1967.
- 46. J. Werner, The HDSL environment (high bit rate digital subscriber line), *IEEE J. Select. Areas Commun.* 9(6): 785-800 (Aug. 1991).
- 47. K. E. Wilson, Analysis of the Crichlow Graphical Model of Atmospheric Radio Noise at Very Low Frequencies M.S. Thesis, Air Force Institute of Technology, Nov. 1974.
- J. W. Wozencraft and I. M. Jacobs, Principles of Communication Engineering, John Wiley & Sons, New York, 1965.
- S. M. Zabin and H. V. Poor, Recursive algorithm for identification of impulsive noise channels, *IEEE Trans. Inform. Theory* 36(3): (May 1990).

STATISTICAL MULTIPLEXING

KAVITHA CHANDRA Center for Advanced Computation and Telecommunications University of Massachusetts at Lowell Lowell, Massachusetts

1. INTRODUCTION

Twenty-first century communications will be dominated by intelligent high-speed information networks. Ubiquitous access to the network through wired and wireless technology, adaptable network systems, and broadband transmission capacities will enable networks to integrate and transmit media-rich services within specified standards of delivery. This vision has driven the standardization and evolution of broadband integrated services network (B-ISDN) technology since the early 1990s. The narrowband ISDN proposal for integrating voice, data, and video on the telephone line was the precursor to broadband services and asynchronous transport mode (ATM) technology. The ISDN and B-ISDN recommendations are put forth in a series of documents published by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T), which was formerly CCITT [1,2] and the ATM Forum study groups [3,4]. The concept of multiplexing integrated services traffic on a common channel for efficient utilization of the transmission link capacity is central in the design of ISDN and ATM networks. ATM in particular advocates an asynchronous allocation of time slots in a time-division multiplexed frame for servicing the variable-bit-rate (VBR) traffic generated from video and data services. ATM multiplexing relies on the transport of information using fixed-size cells of 53 bytes in length and the application of fast cell-switching architectures made possible by advances in digital technology. It utilizes the concepts of both circuit and packet switching by creating virtual circuits that carry VBR streams generated by multiplexing ATM cells from voice, video, and data sources.

The ATM architecture is designed to efficiently transport traffic sources that alternate between bursts of transmission activity and periods of no activity. It also supports traffic sources with continuously changing transmission rates. One measure of traffic burstiness is the ratio of peak to average rate of the source. A circuit-switched network would conservatively allocate to each source a capacity equal to its peak rate. In this case, full resource utilization takes place only when all of the sources transmit at their peak rates. This is typically a low-probability event when the sources are statistically independent of each other. A statistical multiplexer, however, allocates a capacity that lies between the average and peak rates and buffers the traffic during periods when demand exceeds channel capacity. The process of buffering the multiplexed stream smooths the relatively high variations in the traffic rate of individual sources. The multiplexed traffic is expected to have a smaller variance about the mean rate in the limit as the number of sources multiplexed increase to a large value. As a result, there is a diminishing magnitude in the probability of occurrence of source rates that are greater than the available capacity. This feature leads to the economies of scale paradigm of statistical multiplexing that is at the core of B-ISDN and ATM transmission technologies.

Statistical multiplexers have been integral components in packet switches and routers on data networks since the 1960s. They have gained increased prominence since 1990 with the availability of broadband transmission speeds exceeding 155 Mbps and ranging upto 10 Gbps in the core of the network. In conjunction with gigabit switching speeds, the new-generation of internetworks have the hardware infrastructure for delivering broadband services. However, since broadband traffic features are highly unpredictable, the control of service quality such as packet delay and loss probabilities must be managed by a suite of intelligent and adaptable protocols. The development of these techniques is the present focus of standards bodies, researchers, and developers in industry. New Internet protocols and services are currently being proposed by the Internet Engineering Task Force (IETF) to enable integrated access and controlled delivery of multimedia services on the existing Internet packet-switching architecture [5,6]. These include integrated (Intserv) and differentiated (Diffserv) services [7], multiprotocol label switching (MPLS) [8], and resource reservation protocols (RSVPs) [9]. It is expected that ATM and the Internet will coexist with ATM infrastructure deployed at the corporate, enterprise, and private network levels. The Internet will continue to serve connectivity on the wide-area network scale. The design and performance of these new protocols and services will depend on the traffic patterns of voice, video, and data sources and their influence on queues in statistical multiplexers. These problems have been the focus of numerous studies since 1990. This article is organized as follows. Section 2 describes stochastic traffic descriptors and models that have been applied to characterize voice, video, and data traffic. In Section 3 the methods applied for performance analysis of queues driven by the aforementioned models are discussed. Section 4 concludes with a discussion on the open problems in this area.

2. TRAFFIC DESCRIPTORS AND MODELS

The characterization of network traffic with parametric models is a basic requirement for engineering communications networks. Statistical multiplexers in particular are modeled as queueing systems with finite buffer space, served by one or more transmission links of fixed or varying capacity. The service structure typically admits packets of multiple sources on a first-come first-serve (FCFS) basis. Priority-based service may also be implemented in ATM networks and more recent invocations of the Internet protocol. The statistical multiplexing gain (SMG) is an important performance metric that quantifies the multiplexing efficiency. The SMG may be calculated as the ratio of the number of VBR sources that can be multiplexed on a fixed capacity link under a specified delay or loss constraint and the number of sources that can be supported on the basis of peak rate allocation. To determine and maximize the SMG, admission control rules are formulated that can relate traffic characteristics to performance constraints and system parameters.

In this section, the analytic, computational, and empirical approaches for modeling traffic are discussed. A more detailed taxonomy of traffic models is presented by Frost and Melamed [10] and by Jagerman et al. [11]. The traffic is assumed to be composed of discrete units referred to as packets. The packets arriving at the multiplexer input are characterized using the sequence of random arrival times T_1, T_2, \ldots, T_n measured from an origin assumed to be zero. The packets are associated with workloads W_1, W_2, \ldots, W_n that may also be random variables. These workloads can represent variable Internet packet sizes fixed ATM cell sizes, or in case of batch arrivals, where more than one packet may arrive at a time instant, the workload represents the batch size. The packet interarrival times $\tau_n = T_n - T_{n-1}$ or the counting process N(t), which represents the number of packets arriving in the interval (0, t] are representative and equivalent descriptors of the traffic.

The most tractable traffic models result when interarrival times and workload sequences are independent random variables and independent of each other. A renewal process model is readily applicable as a traffic model in such a case. Telephone traffic on circuit-switched networks has been shown to be adequately modeled by independent negative exponential distributions for the interarrival times and call holding times. As shown by A. K. Erlang in his seminal study [12] of circuit-switched telephone traffic, the Poisson characteristics of teletraffic greatly simplify the analysis of queueing performance. Packet traffic measurement studies since 1970 have, however, shown that the arrival process of data, voice, and video applications rarely exhibit temporal independence. Traffic studies [13-15] conducted during the Arpanet days examined data traffic generated by user dialogs with distributed computer systems and showed that computer terminals transmitted information in bursts that occurred at random time intervals. Pawlita [16] presented a study of four different user applications in data networks and identified bursty traffic patterns, clustered dialog sequences and hyperexponential distributions for the user dialog times. Traffic measurement studies conducted on localarea networks [17-19] and wide-area networks [20] have found similar statistics in the packet interarrival times.

A measurement and modeling study of traffic on a token ring network by Jain and Routhier [17] showed

that the packet arrivals occurred in clusters, for which they proposed a packet train model. The time between packet clusters was found to be a function of user access times, whereas the intracluster statistics were a function of the network hardware and software. More recent analyses of Internet traffic by Paxson and Floyd [21] and Caceres et al. [22] have shown that packet interarrival times generated by protocol-based applications such as file transfer, network news protocol, simple mail transfer, or remote logins are neither independent nor are they exponentially distributed.

Meier-Hellstern et al. [23], in their study of ISDN data traffic, have shown that the interarrival times for a user's terminal generated packet traffic can be modeled by superposing a gamma and power-law type probability density functions. The traffic generated in an Ethernet local area network of workstations has been shown by Gusella [24] to be nonstationary and characterized by a long-tailed interarrival time distribution. Leland et al. [25] analyzed aggregated Ethernet traffic on several timescales. A self-similar process was proposed as a model based on scale invariant features in the traffic. This model implies that the traffic variations are statistically similar over many, theoretically infinite ranges of timescales. As a result, one observes temporal dependence in the traffic structures over large time intervals. Erramilli et al. [26] propose a deterministic model based on chaotic maps for modeling these long-range dependence features. A compilation of references to work done on selfsimilar traffic modeling can be found in the study by Willinger et al. [27]. The aforementioned data traffic studies indicate that temporal dependence features found in measurements must be described accurately by traffic models. In this regard, static traffic descriptors such as the first- and second-order moments and marginal distributions of the traffic have been proposed. Dynamic models that capture some of the temporal features of the arrival process have also been proposed.

Traffic bursts are structures characterized by a successive occurrence of several short interarrival times followed by a relatively long interarrival time. This feature has been characterized using simple first-order descriptors such as the ratio of peak to average rate. In terms of the random interarrival times τ , the coefficient of variation c_{τ} captures the dispersion in the traffic through the ratio of the standard deviation and the expectation of the interarrival times

$$c_{\tau} = \frac{\sigma[\tau]}{E[\tau]} \tag{1}$$

Alternately, the index of dispersion $I_N(t)$ of the counting process N(t) can be calculated for increasing time intervals of length t. This is a second-order characterization that captures the burstiness as a function of the variance of the process and is given by

$$I_N(t) = \frac{\text{Var}\left[N(t)\right]}{E[N(t)]} \tag{2}$$

An index of dispersion for intervals $I_{\tau}(n)$ may be similarly defined by replacing the numerator and denominator of Eq. (2) by the variance and expectation of the sum

2422 STATISTICAL MULTIPLEXING

of n successive interarrival times. The correlation in the workloads may also be characterized using the aforementioned indices of dispersion. The magnitude and rate of increase in these traffic descriptors can capture succintly, the degree of correlation in the arrival process. An increasing magnitude of the index of dispersion with the observation time indicates highly correlated streams that are in turn linked to large packet delays and packet losses. For example, the expected number in a single server queue driven by Poisson arrivals and a general service time distribution (M/G/1) is given by the Pollaczek-Khintchine mean value formula [28], which shows that the average queue size increases in direct proportion to the square coefficient of variation of the service times. For stationary arrival processes, the limiting values of the indices of dispersion as n and t tend to infinity are shown [24] to be related to the normalized autocorrelation coefficients $\rho_{\tau}(j) \ j = 0, 1, 2, \dots, \text{ as}$

$$I_N = I_{\tau} = c_{\tau}^2 \left[1 + 2 \sum_{j=1}^{\infty} \rho_{\tau}(j) \right]$$
(3)

Sriram and Whitt [29] and others [30] apply the index of dispersion of counts (IDC) and intervals for examining the burstiness effects of superposed packet voice traffic on queues. The IDC of a single packet voice source approached a limiting value of 18 in comparison to a value of unity for a Poisson process. It has been shown [29] that under superposition, the magnitudes of the IDC of the multiplexed process approached Poisson characteristics for short time intervals. As the time interval increased the positive autocorrelations of the individual sources interact, leading to increased values of the IDC parameter. The larger the number of sources superposed, the larger is the time interval at which the superposed process deviates from Poisson-like statistics. These concepts showed the importance of identifying a relevant timescale for the superposed traffic that allows the sizing of the buffers in a queue. Although the index of dispersion descriptors have proved useful for evaluating the burstiness property in a qualitative way, they have limited application for deriving explicit measures of queue performance.

A method for estimating an index of dispersion of the queue size using the peakedness functional is presented by Eckberg [31,32]. The "peakedness" of the queue represents the ratio of the variance and expectation of the number of busy servers in an infinite server system driven by a stationary traffic process. This approach incorporates second-order traffic descriptors. The traffic workloads represented by random service times S are modeled by the service time distribution F(t), its complement Q(t) = 1 - 1 $F(t) = \Pr[S > t]$, and the autocorrelation of Q(t) denoted by $R_Q(t) = \int_0^\infty Q(x)Q(t+x) dx$. In addition, the arrival process may be characterized by a time-varying, possibly random, arrival rate $\lambda(t)$ and its covariance density $k(\tau)$. An arrival process characterized by a random arrival rate belongs to the category of doubly stochastic processes. Cox and Lewis [33] derive the covariance density of a doubly stochastic arrival process as $k(\tau) = \sigma_{\lambda}^2 \rho_{\lambda}(\tau)$, where σ_{λ}^2 and $\rho_{\lambda}(\tau)$ are the variance and normalized autocovariance

functions of $\lambda(t)$, respectively. With the traffic specified by these functions, the expected value and variance of the number of busy servers L(t) at a time t may be obtained as follows:

$$E[L(t)] = \int Q(t-\tau)\lambda(\tau) d\tau \qquad (4)$$

$$Var [L(t)] = \int [Q(t-\tau)\{1-Q(t-\tau)\}\lambda(\tau) + k(\tau)R_{Q}(\tau)] d\tau \qquad (5)$$

The presence of correlations in the arrival rate for lags greater than zero cause the arrival process to be overdispersed relative to Poisson processes with constant arrival rate. The degree of dispersion is proportional to the magnitude of $k(\tau), \tau > 0$ and the decay rate of Q(t). This increased traffic variability has an impact on the problem of resource allocation and engineering. The peakedness functional $Z[F] = \frac{E[L(t)]}{\operatorname{Var}[L(t)]}$ provides a measure of the influence of traffic variance and correlations on queue performance. Z[F] has a magnitude that is greater than one for processes with nonzero k(s), s > 0. The peakedness of the process influences the traffic engineering rules used for sizing system resources. The application of peakedness to estimate the blocking probability of finite server systems is presented by Fredericks [34]. Here a knowledge of Z[F] is used in Hayward's approximation, an extension of Erlang's blocking formula to estimate the additional servers needed for Z[F] > 1. The utility of the peakedness characterization for analysis of delay systems has been discussed by Eckberg [32].

A more descriptive representation of the multiplexer queues is provided by the steady-state probability distribution of the buffer occupancy. If the random variable X represents the buffer occupancy in the steady state, the shape of the complementary queue distribution $G(x) = \Pr[X > x]$ provides information on the timescales at which traffic burstiness and correlations impact the queue. Livny et al. [35] have shown that the positive autocorrelations in traffic have significant impact in generating increased queue sizes and blocking probabilities relative to independent identically distributed processes. In this context it is useful to differentiate between two types of queue phenomena: queues arising from packet or cell level congestion and those arising from burst level congestion [36]. Packet level queues occur due to an instantaneous arrival of packets from different sources in the same time-slot resulting in a cumulative rate that is greater than the service capacity. It may be due to the chance occurrence of a set of interarrival times of different sources that cause individual packets to collide in time at the multiplexer input. This phenomenon can also occur for deterministic traffic, such as periodic sources, when the starting epochs are randomly displaced from one another [37,38]. Queues arising from packet-level congestion are typically of small to moderate size and can be accommodated using small buffers.

Larger queue sizes result when multiple traffic sources start transmitting in the burst state. Here, sustained transmission of a number of sources at the peak rate leads to a buildup in the queue size for time durations that are functions of the burst state statistics. Since individual times of packet arrivals are not important in this case, burst-level queues have been analyzed using the fluid flow model [39]. In the fluid approximation, the discrete packet arrival process and the buffer occupancy variables are replaced by real-valued random processes. Although burst level congestion leads to lower probability events than does packet level congestion, the decay rate of these probabilities are functions of both the service rate and traffic source burst statistics. Figure 1 shows a depiction of a typical structure of the packet- and burst-level queue components in G(x). In the design of a statistical multiplexer the size of buffer is typically set to absorb packet-level queues. Burst-level queues estimated from infinite buffer queue analysis can be used to approximate the losses that take place in a finite buffer system. Finite buffer systems typically require more complex analysis than infinite buffer systems.

The fluid approximation requires that the time variation and correlation of the arrival rate process be prescribed. Finite-state Markov chain models of traffic have been applied extensively in fluid buffer analysis. Discrete- and continuous-time Markov chains (DTMC, CTMC) with finite-state space [40] are among the simplest extensions to the renewal process model for incorporating temporal dependence. The traffic correlation structure exhibits geometric or exponential rate of decay for the discrete and continuous-time Markov chains, respectively. A *K*-state discrete time Markov chain Y[n], n = 0, 1, 2, ...resides in one of K states S_1, S_2, \ldots, S_K at any given time n. By the Markov property, the probability of transitioning to a particular state at time n is a function of the state of the process at n-1 only. These one-step transition probabilities are specified in a K-dimensional matrix P_Y as elements $p_{ij} = \Pr[Y[n] \in S_j | Y[n-1] \in S_i]$. The elements in each row of P_Y sum to unity. In a continuous time Markov chain, the transition rates are captured by an infinitesimal generating matrix Q_Y containing elements



Figure 1. Packet- and burst-level components of queue size distribution.

 q_{ij} that represent the transition rate from state S_i to S_i for $i \neq j$. In this case, the sum of the rates in each row is equal to zero. The probability transition matrix and the generator matrix uniquely determine the rate of decay in the autocorrelations of the Markov chain. In the context of modeling traffic arrivals the transition matrix is supported by a K-state rate vector that describes the arrival rate when the traffic is in a particular state. This feature allows the variable rate features of network traffic to be represented. The rate vector in the simplest case is a set of constants that may represent the average traffic rate in each state. More general models based on a stochastic representation for the rate selection have also been considered. The Markov modulated Poisson process (MMPP) [30,41,42] is one example where the Markov process is characterized by a state-dependent Poisson process. These Markovian models of traffic can capture time variations in the arrival rate and associate these variations with a temporal correlation envelope that is determined by the magnitude of the transition probabilities. These models, however, cannot address nonexponential trends in the correlation function. To accommodate more general shapes of the correlation functions, Li and Hwang [43,44] propose the application of linear systems analysis using power spectral representation of traffic.

3. PERFORMANCE OF STATISTICAL MULTIPLEXERS

Statistical multiplexing is designed to increase utilization of a resource that is subject to random usage patterns. In this work, the resource is considered to be a transmission link of finite capacity. Multiple sources access the channel on a first-come first-serve or priority basis and are allowed to queue in a buffer when the channel is busy. Statistical multiplexing gains come at the expense of a packet loss or delay probability that is considered tolerable for the applications being transported. The performance constraint is typically specified by the acceptable probability of loss for a given buffer size B. In some cases an infinite buffer queue is analyzed for tractability and the loss probability is approximated by the tail probability of queue lengths P(X > B). For the limiting case of zero buffer size, the probability of loss may be calculated by determining the probability of the aggregate input rate exceeding the capacity.

The approaches to performance analysis in the literature may be classified by applications. The multiplexing of packet voice with data using two state Markov chains to model the ON and OFF states and the application of analytic and simulation-based performance analysis was the subject of numerous studies since the 1970s. With the availability of larger transmission capacities and standardization of encoding schemes for digital video, the transport and multiplexing of packet video became an active area of research in the early 1990s. Models for variable-bit-rate (VBR) video were found to be more complex and of higher dimensionality. The Markov representation for packet video typically required a large state space to capture the temporal variations and amplitude distributions. As a result, several approximation methodologies such as effective bandwidth formalisms and large

deviations analyses were proposed to relate the traffic characteristics, performance constraints, and statistical multiplexer parameters. These approximations have been particularly useful in formulating admission control decisions for multiservice networks.

In the following section a review of voice/data multiplexing schemes is provided first. This is followed by a presentation of video traffic models and the analysis of their multiplexing performance using fluid buffer approximations. This leads to a discussion of admission control algorithms with particular focus on the effective bandwidth approximations.

3.1. Voice and Data Multiplexers

In the early 1980s integrated services digital networks (ISDNs) [45] were envisoned to support multiplexed transport of voice-, data-, and image-based applications on a common transport infrastructure that included both telephone and data networks. The digital telephone network with a basic transmission rate of 64 kbps (kilobits per second) was considered to be the dominant transport network. Different local user interfaces were standardized to connect end systems such as telephones, data terminals, or local area networks to a common ISDN channel. At any given time, the traffic generated on this link could be a mix of data, voice, and associated signaling and control information. The performance requirements of voice and data traffic [46] govern the design of the multiplexing system. The buffer overflow probability is a chief concern for data transmission, whereas bounding transmission delay is critical for speech signals [47]. Initial studies on the performance of voice/data multiplexing systems assumed fixed duration time-division multiplexing frames in which time slots were distributed between voice and data packets. In this context, the multiplexing efficiency of voice and data has been analyzed using various approaches that involve moving frame boundaries between voice and data slots [48-50], separate queueing buffers for voice and data [51], encoder control for voice [52], application of circuit-switching concepts for both voice and data [53], and hybrid models of circuitswitched voice/packet-switched data [54]. Maglaris and Schwartz [55] describe a variable-frame multiplexer that admits long messages of variable length and single packets that arrive as a Poisson process. The Poisson model is assumed to impose a degree of traffic burstiness on the otherwise continuous rate process. The ability to adapt frame sizes in response to traffic variations showed improved performance in terms of bandwidth utilization and delays relative to fixed-frame movable-boundary schemes. The system requirements for multiplexing data during silence periods of speech is presented by Roberge and Adoul [56]. In this work the accurate discrimination of speech and data signals is proposed using statistical pattern classification algorithms based on zero-crossing statistics of the quadrature-amplitude-modulated speech signal. A speech-data transition detector is proposed for detecting switching points in time with accuracy.

With the evolution of fast packet switching devices, the more recent approaches to voice/data integration have examined the performance of asynchronous multiplexing on a single high-speed channel. Voice/data integration concepts were motivated in part, by the traffic characteristics of data applications such as Telnet, File Transfer Protocol (FTP), and Simple Mail Transfer Protocol (SMTP). These applications generate bursts of activity separated by random durations of inactive periods. Speech patterns in telephone conversations are also characterized by random durations of talk spurts that are followed by silence periods. Brady [57] presented experimental measurements of the average durations of ON and OFF periods and transition rates between these states from a study of telephone conversations. Typically the average speech activity is found to range from 28 to 40% of the total connection time and is a function of users, language, and other such factors. Average length of talk and silence spurts are in the 0.4-1.2-s and 0.6-1.8-s ranges.

A two-state Markov process [58] representing the ON and OFF states has been the canonical model for characterizing speech-based applications, although the silence durations are seldom exponentially distributed. The characteristic transition rates between ON and OFF states can be significantly different for voice and data sources. The alternating talk spurts and silence durations of speech applications exhibit relatively slow transition rates, allowing the data to be multiplexed in the OFF periods. Data sources exhibit faster transitions between active and inactive states. A problematic feature in packet voice traffic is its temporal correlation which is induced by speech encoders and voice activity detectors [29,30]. As a result of multiplexing with voice in a queue, the departing data flow takes on the characteristics of the superposed voicestream. This feature influences the performance of other multiplexers in the transmission path.

Heffes and Lucantoni [30] model the dependence features of multiplexed voice and data traffic using a two-state Markov modulated Poisson process (MMPP). Asynchronous voice data multiplexing of MMPP sources is examined by evaluating the delay distributions of a single-server queue with first-in first-out (FIFO) service and general service time distribution. The application of this model for evaluation of overload control algorithms is discussed. Sriram and Whitt [29] extract the dependence features of aggregate voice packet arrival process from a highly variable renewal process model of a single voice source. The aggregation of multiple independent voice sources is examined using the index of dispersion of intervals (IDI). The motivation behind this approach is that the limiting value of IDI as number of sources tend to infinity completely characterizes the effect of the arrival process on the congestion characteristics of a FIFO queue in heavy traffic. This work also shows that the positive dependence in the packet arrival process is a major cause of congestion in the multiplexer queue at heavy loads. Buffer sizes larger than a critical value as determined by the characteristic correlation time scale will allow a sequence of dependent interarrival times to build up the queue, causing congestion. Limiting the size of the buffer, at the cost of increased packet loss is proposed as an approach for controlling congestion. To control packet loss that occurs from dependence in arrival process, Sriram and Lucantoni [59] propose

dropping the least significant bits in the queue when the queue length reaches a given threshold. They show that under this approach the queue performance is comparable to that of a Poisson traffic source. These pioneering studies provided a comprehensive understanding on the efficiency of synchronous and asynchronous approaches for multiplexing voice and data traffic on a common channel. A quantitative characterization of the dependence features in traffic was shown to be one of the most important requirements for performance evaluation. In this regard, finite-state Markov processes were found to be amenable in both capturing some of the dependence features and allowing tractable analysis of the multiplexer queues. These studies also had limitations in that traffic measurements and measurement based models did not play a prominent role in analysis of multiplexers. However, with transition from ISDN to B-ISDN and the recognition that simple two-state Markov models are inadequate for broadband sources, more emphasis has been placed on measurement based analysis of video and data traffic. The developments in video models and multiplexers are discussed next.

3.2. Video Models and Multiplexers

Video communication services are important bandwidth consuming applications for B-ISDN. In an early study, Haskell [60] showed that multiplexing outputs of picturephone video encoders into a common buffer could achieve significant multiplexing gains. Although current compression techniques for digital video can achieve video bit rates of acceptable quality in the range of 1-5 Mbps, when hundreds of such flows are to be transported, efficient multiplexing schemes are still required. Figure 2 depicts a comparison of the temporal variation of measured video frame rates for a low-activity videoconference encoded with H.261 standard and high-activity MPEG-2 encoded entertainment video. The signals represent the number of bits in each encoded video frame as a function of the frame index. The large dispersions about a mean rate are evident, as are the sudden transitions in frame rate amplitudes when encoding changes from predictive to refresh mode.

The transport of variable bit rate (VBR) video using statistical multiplexing has been examined in numerous studies. VBR video is preferred over traditional constant-bit-rate (CBR) video due to the improved image quality and shorter delays at the encoder. Statistical multiplexing invariably results in buffering delays and losses, which can significantly degrade video quality. To minimize the amount of delay and loss, the networking community has focused on the development of effective and implementable congestion control schemes, including connection admission control and usage parameter control. To minimize the impact of delay and loss, the video community has focused on developing good error concealment algorithms and designing efficient two-layer coding algorithms [61,62] for use in combination with the dual-priority transport provided by ATM networks. For example, while one-layer MPEG-2 produces generally unacceptable video quality with a cell loss ratio of 10^{-3} , losses at this rate with SNR scalability (one of the four standardized layered coding algorithms of MPEG-2) are generally invisible, even to experienced viewers [63].

Various application- and coding-specific models of onelayer VBR video have been proposed in the literature. Maglaris et al. [64] was among the first to analyze short (10-s) segments of low-activity videophone signals. A first



Figure 2. Sample paths of VBR video in videoconferencing and entertainment applications.

order autoregressive (AR) process was proposed for the number of bits in successive video frames of a single source. The multiplexed video was modeled by a birth-death Markov process. In this model, transitions are limited to neighboring states. The model parameters were selected to match the mean and short-term covariance structure in the measurements. The states of the Markov chain were derived by quantizing the aggregate source rate histogram into a fixed number of levels. A choice of 20 levels per source was found adequate for the lowactivity videoconferencing source. Sen et al. [65] extended this model to accommodate moderate activity sources, using additional states to model low and high activity levels. The resulting source model is equivalent to that obtained by superposition of independent ON-OFF processes. Grunenfelder et al. [66] used a conditional replenishment encoder that exhibits strong correlations effects. The superposed video process was assumed to be wide sense stationary. The multiplexer was modeled using a general arrival process with independent arrivals and deterministic service times. A source model for full motion video was presented by Yegenoglu et al. [67] using an autoregressive process with time-varying coefficients. The selection of the coefficients was based on the state of a discrete-time Markov chain. The transition and rate matrices were constructed by matching the rate probability density function with that obtained from measurements. Moderate-activity videoconference data were also modeled by Heyman et al. [68]. The rate evolution was modeled by a Markov chain with identical transition probabilities in each row. These probabilities were modeled as a negative binomial distribution. The number of states in the model was of the order of the peak rate scaled by a factor of 10. This ranged from 400 to 500 states. The within state correlations were modeled by a discrete AR (DAR) process that resulted in a diagonally dominant matrix structure. This structure did not model single-source statistics very well since there were no selective transitions between states based on source characteristics. The DAR model failed to capture the short-term correlations in the traffic of a single source. Lucantoni et al. [69] proposed a Markov renewal process (MRP) model for VBR source traffic. Results of this model show that burstiness in video data can be captured more accurately than in the DAR model. Although the MRP was shown to perform better than the DAR model in capturing the burstiness, the MRP still did not match the cell loss probabilities for large buffer sizes. Heyman and Lakshman [70] studied high-activity video sources and concluded that their DAR model proposed for videoconference sources could not be applied as a general model to all sources. Skelly et al. [71] also used Markov chains to verify a histogram-based queueing model for multiplexing. They determined, on the basis of simulation, that a fixed number of eight states were sufficient to model the video source.

The video encoding system effects play a significant role in shaping the temporal and amplitude variation of compressed video. Traffic shaping algorithms at policing systems that enforce constraints on the output rate of the encoding system also play an important role in shaping digital video traffic [72-74]. In general, the coder that produces a bit stream conforming to constraints will not have the same statistical characteristics of an unconstrained coder. The idea that encoders could be constrained to generate traffic described by Markov chains was explored by Heeke [75] for designing better traffic policing and control algorithms. Pancha and Zarki [76] have examined the traffic characteristics resulting from various combinations of the quantization parameter, the inter-to-intraframe ratio and the priority breakpoint in MPEG one-layer and two-layer encoding, respectively. Data generated for each parameter set were modeled by a Markov chain by selecting the number of states based on the ratio of peak rate to the standard deviation of the frame rates. Frater et al. [77] verify the performance of a non-Markovian model for full motion video based on scene characterization by matching the cell loss probabilities at different buffer sizes. Krunz and Hughes [78] modeled MPEG, with distinct models for each different frame type. The selection of an adequate number of states in the Markov chain model of video such as to adequately model the spectral content is discussed by Chandra and Reibman [79]. Adequate spectral content in single sources is found necessary to understand the scaling aspects of Markov models under multiplexing.

Non-Markovian models exhibiting long-range dependence have been proposed by Garrett and Willinger [80] and others [77,81]. Ryu and Elwalid [82] show that longterm correlations do not significantly affect network performance over a reasonable range of cell losses, buffer sizes, and network operating parameters. Grossglauser and Bolot [83] also propose that correlation timescales to be considered in the traffic depend on the operating parameters and that full long-range dependence characterization of traffic is unnecessary. The impact of temporal correlation in the output rate of a VBR video source on the queue response has been examined [43], and it has been shown that macrolevel correlations can be modeled by Markov-chain-based models. Long-range dependence seen in VBR video has also been examined and the queueing results have been compared to those obtained using the DAR model [84]. It was concluded that for moderate buffer sizes, the short-range correlations obtained using Markov chain models are sufficient to estimate the buffer characteristics.

The aforementioned studies have determined that correlations on many timescales are an inherent feature in video sources and that Markov modulated source models are appropriate for capturing these dynamics. The ubiquitous use of multistate Markov models has led to work on their performance in queues. The application of finite-state Markov chain video traffic models for H.261 and MPEG coders in simulation studies [79,85] has shown that with an increase in the number of multiplexed video sources and corresponding increase in the channel capacity, the loss probability can be significantly reduced and reasonable multiplexing gains achieved. It was shown that typically 15-20 states are required to faithfully model the queue behavior of moderate to high-activity video sources. When using too few states, the tail probabilities of the rate histogram will not be captured, thereby yielding

an underestimate of the packet delay or loss probability. This situation has been observed by Hasslinger [86] in modeling VBR sources using semi-Markov models.

The performance analysis of queues driven by largedimensional Markovian traffic sources may be approached using exact queueing analysis in discrete time and discrete state space. This approach becomes quickly intractable as the number of sources increase due to exponential increase in state dimension. In the limit of a large number of sources operating in the heavy-traffic regime, the discrete arrival and departure times may be replaced by a fluid approximation. This analysis technique is discussed in the next section.

3.3. Fluid Buffer Models

Fluid flow models assume that the packet arrival process at a multiplexer occurs continuously in time and may be characterized by continuous random fluctuations in the arrival rate [87]. This approach is applicable when the packet sizes are small relative to the link capacity. The computational model presented by Anick et al. [39] affords the estimation of the delay and loss distributions in multiplexers fed by Markov modulated fluid sources and served at constant rate. In this method, the buffer occupancy X is assumed to be a continuous valued random variable. The arrival process of each source is represented by a finite-state continuous-time Markov generator Q and associated diagonal rate matrix R. If K is the number of states required to represent a single source, and N is the number of sources (assumed identical) being multiplexed, the superposition can be modeled by the Markov generator Q_N and diagonal rate matrix R_N , which are computed as the N-fold Krönecker sums $Q \oplus Q \oplus \cdots \oplus Q$ and $R \oplus R \cdots \oplus R$, respectively. The Krönecker sum operation increases the dimension of the multiplexed source generator matrix to $M = K^N$.

The aggregated traffic stream enters a queue with finite or infinite waiting room. Packets in the buffer are serviced on a first-in first-out basis at a constant service rate. The cumulative probability distribution of the buffer occupancy x in steady state is specified by the row vector $\vec{p}: [p_0(x), p_1(x), \ldots, p_{M-1}(x)]$, where element $p_i(x) = \text{Prob} [X \leq x; source in state i]$. For a service rate of C packets per second, the probabilities satisfy the equation

$$\frac{\partial \vec{p}}{\partial x}D = \vec{p}Q_N \tag{6}$$

where the matrix $D = [R_N - IC]$ captures the drift from the service rate in each state. Here *I* is the identity matrix. The solution of Eq. (6) follows that of an eigenvalue problem and may be represented in terms of the tail probability distribution G(x) as

$$G(x) = \Pr[X > x] = \sum_{i=0}^{M-1} a_i(x)e^{-z_i x}$$
(7)

where z_i , i = 0, ..., M - 1 are the eigenvalues of the matrix $Q_N D^{-1}$. The coefficients $a_i(x)$ are functions of the eigenvalues and eigenvectors [88,89] of $Q_N D^{-1}$. For an infinite buffer, subject to consideration that the solution

is bounded at x equal to infinity, the coefficients of the exponentially growing modes are set equal to zero. The amplitudes of the remaining modes are determined by applying the appropriate boundary conditions for overload and underload states. Underload states represented by states of the drift matrix with negative elements are subject to the condition $p_i(x = 0) = 0$. The coefficients for overload states are solved by equating $p_i(\infty)$ to the steady-state probability of the multiplexed source being in state *i*. For a buffer of finite size, all of the eigenvalues are retained and boundary condition at infinity replaced by the corresponding value at the buffer size.

The aforementioned approach requires the solution of an eigenvalue problem for a matrix whose upper bound on dimension scales as $O(K^N)$. To counter this dimensionality problem, reduced order traffic models have been used as approximations. For two-state ON-OFFMarkov processes the superposition yields a generator of O(N) states. Multi-state Markov sources have been approximated by the superposition of multiple two-state ON-OFF sources by matching first and second moments of the two processes [64,65,90]. The number of two-state sources selected for this model is often an arbitrarily choice. For moderate to high-activity video sources, this approximation can be shown to underestimate the packet delays.

Correlation effects afforded by the generator matrix play a dominant role in structuring the features of burstlevel queueing delays. A traffic source represented by a finite-state Markov chain exhibits temporal autocorrelations that decay exponentially in time. The rate of decay is governed by the dominant eigenvalues of the generator matrix Q. It can be shown that the characteristic correlation time-scale of a single source is retained in the superposed traffic. The selection of an adequate single source model order K that captures all of the dominant modes is therefore an important consideration in building the traffic model. High-activity video sources often require K to be in the range of 15-20 states. Figure 3 shows the effect of choosing an inadequate number of states K for modeling the H.261 encoded videoconference source shown in Fig. 2a. As K is increased from 5 to 16, the asymptotic decay rate of the complementary delay distribution approaches that exhibited by the measurements.

For sources with large-dimensional K, even for moderate values of N, the estimation of buffer occupancy distributions for the multiplexed system becomes computationally intensive. A method for reducing the state-space dimension of multiplexed source generator is given by Thompson et al. [91]. The reduction process involved the quantization of the rates and the fundamental rate was chosen to yield the best match to the mean and variance of the rate. States having equal rates were aggregated, thereby reducing the number of states in the generator matrix. The resulting model allowed for scalable analysis as the number of sources was increased.

For large-dimensional systems, asymptotic approximations to the model given in Eq. (7) may be obtained for large buffer sizes and small delay or loss probabilities [92]. In this approximation

$$G(x) \sim \alpha e^{-\beta x} \quad x \to \infty$$
 (8)



Figure 3. Influence of number of states *K* chosen to model video source.

where α is referred to as the asymptotic constant and β is the largest negative eigenvalue of the matrix $Q_N D^{-1}$. The asymptotic decay rate $-\beta$ is a function of the service rate C and may be determined with relative ease. The asymptotic constant, however, requires knowledge of all the eigenvectors and eigenvalues of the system. Approximate methods for estimating α for Markovian systems have been derived [93,94]. Although most of the asymptotic representations have considered Markovian sources, there have been some results for traffic modeled as stationary Gaussian processes [95] and fractional Brownian motion [96].

Very often, $\alpha = G(0)$ is assumed to be unity in the heavy-traffic regime. This allows a very usable descriptor of multiplexer performance that is referred to as the *effective bandwidth* of a source, which assumes the limiting form of the tail probabilities be structured as

$$G(x) \approx e^{-\beta x} \tag{9}$$

To achieve a specified value of β that satisfies given performance constraints, the required capacity may be shown [97] to be obtained as the maximal eigenvalue of the matrix $[R_N + \frac{Q_N}{\beta}]$. This capacity is referred to as the effective bandwidth (EB) of the multiplexed source. In the limit as β approaches zero and ∞ , EB approaches the source average and peak rate, respectively. However, as noted by Choudhury et al. [98], the EB approximation can lead to conservative estimates for bursty traffic sources that undergo significant smoothing under multiplexing. It was shown that the asymptotic constant was itself asymptotically exponential in the number of multiplexed sources N. For traffic sources with indices of dispersion greater than Poisson, this parameter decreased exponentially in N, reflecting the multiplexing gain of the system. The application of these approximations in designing efficient multiplexing systems through admission control is discussed next.

3.4. Effective Bandwidths and Admission Control

Network traffic measurements have identified application and system dependent features that influence the traffic characteristics. Characterization in terms of the mean traffic rate is inadequate because of the large variability in its value over time. As traffic mixes and their performance requirements change, network mechanisms that adapt to these variations are of critical importance in broadband networks. Admission and usage control policies are two proposals in place in ATM networks and the next-generation Internet An admission control process determines if a source requesting connection can be admitted into the system without perturbing the performance of existing connections. This control algorithm should therefore make its decision taking into account a specific set of traffic parameters, available link capacity, and performance specifications. If a flow is admitted, the usage control policy monitors the flow characteristics to ensure that its bandwidth usage is within the admitted values.

To facilitate admission control, the concept of service classes has been introduced to categorize traffic with disparate traffic and service characteristics. Multiplexing among the same and across different service classes have been analyzed. The performance of five classes of admission control algorithms is reviewed by Knightly and Shroff [99]. These classes include scheduling based on average and peak rate information [100], effective bandwidth calculations [97,101,102], their refinements from large deviation principles [103], and maximum variance approaches based on estimating the upper tails of Gaussian process models of traffic [95]. A overview of the EB and its refinements is presented, since it appears to be the most generally applicable formalism.

It is assumed that K classes of traffic are to be admitted into a node served by a link capacity C packets per second. If N_i sources of type *i* exist, each characterized by an effective bandwidth E_i , then the simplest admission policy is given by the linear control law:

$$\sum_{i=1}^{K} N_i E_i \le C \tag{10}$$

The effective bandwidths are derived taking into consideration the traffic characteristics and performance requirements of each class and available capacity C. Defining the traffic generated by type i source on a timescale t by a random variable $A_i[0, t]$, the effective bandwidth derived from large-deviation principles is given by [103]

$$E_i(s,t) = \frac{1}{st} \log E[e^{sA_i[0,t]}]$$
(11)

where the parameter s is related to the decay rate of G(x)and captures the multiplexing efficiency of the system. It is calculated from the specified probability of loss or delay bounds [89]. The term on the right is the log moment generating function of the arrival process. The workload can be described over a time t that represents the typical time taken for the buffer to overflow starting from an empty state. For a fixed value of t, the EB is an increasing function of s and lies between the mean and peak values of $A_i[0,t]$. This may be shown by a Taylor series approximation of E_i as $s \to 0$ and $s \to \infty$ respectively. Methods for deriving E_i for different traffic classes are discussed by Chang [104]. The aforementioned model assumes that all the multiplexed sources have the same quality of service requirements. If not, all sources achieve the performance of the most stringent source. Kulkarni et al. [105] consider an extension of this approach for addressing traffic of multiple classes. For the superposition, since the total workload is given by $A[0, t] = \sum_{i=1}^{K} A_i[0, t]$ the effective capacity C_i of the

by $A[0,t] = \sum_{i=1}^{k} A_i[0,t]$, the effective capacity C_e of the multiplexed system is

$$C_e = \sum_{i=1}^{K} E_i \tag{12}$$

The admission control algorithm simply compares C_e with available capacity C and if $C_e < C$ allows the new source to be admitted into the system.

4. CONCLUSIONS AND OPEN PROBLEMS

The current status on statistical multiplexing in broadband telecommunication networks has been presented. The multiplexing issues in the early 1980s were concerned with voice and data integration on 63-kbps telephone channels. Variations of synchronous time-division multiplexing using moving boundaries between voice and data slots, silence detection for insertion of data packets, and development of adaptive speech encoders were the primary concerns in designing efficient multiplexers. The transition to broadband era characterized by capacities exceeding 155 Mbps evolved with the design and standardization of asynchronous transfer mode networks. ATM networks were envisioned to integrate and optimize features of circuit- and packet-switched networks. The increase in switching speeds and network capacities in the 1990s and the invention of the World Wide Web concept, accelerated the development of many new applications and services that involved networked voice, video, and data. With increased accessibility of the Internet, many of the ATM related paradigms such as traffic characterization, admission control, and statistical multiplexing efficiency are now more relevant for the public Internet.

The important open issues at present are robust characterization of traffic and the derivation of traffic models that can be tractably analyzed in a queueing system. With the automation of end systems and application of complex encoders and detectors, the traffic patterns seen on networks today may not readily map to a pure stochastic model framework. On the contrary, traffic measurements indicate that deterministic patterns and nonlinearities in traffic amplitudes are new prevailing features in broadband networks. Large-dimensional stochastic models are seen to be required to capture these features. The computational complexity associated in analyzing the multiplexing problem for such models has led to some innovative approximation techniques. The characterization of a traffic source by an effective bandwidth is one such result that is derived by application of large-deviation principles. Derived in the asymptotic limit of large number of multiplexed sources, large buffer sizes and link capacities and small probabilities of delay and loss, effective bandwidths offer a conservative, but computationally feasible model for evaluating multiplexing efficiency in theory. The design of real-time algorithms for applying these concepts on a network and discovery of their effectiveness is expected to be the next step in the statistical multiplexing analysis.

BIOGRAPHY

Kavitha Chandra received her B.S. degree in electrical engineering in 1985 from Bangalore University, India, her M.S. and D.Eng. degrees in computer and electrical engineering from the University of Massachusetts at Lowell in 1987 and 1992, respectively. She joined AT&T Bell Laboratories in 1994 as a member of technical staff in the Teletraffic and Performance Analysis Department. From 1996 to 1998 she was a senior member of technical staff in the Network Design and Performance Analysis department of AT&T Laboratories. She is currently an associate professor in the Department of Electrical and Computer Engineering and principal faculty in the Center for Advanced Computation and Telecommunications at the University of Massachusetts at Lowell. Dr. Chandra received the Eta Kappa Nu Outstanding Electrical Engineer Award (honorable mention) in 1996 and the National Science Foundation Career Award in 1998. Her research interests are in the areas of network traffic and performance analysis, wireless networks, acoustic and electromagnetic wave propagation, adaptive estimation and control.

BIBLIOGRAPHY

- CCITT Red Book, Integrated Services Digital Network (ISDN), Series I Recommendation, Vol. III, Fascicle III.5, 1985.
- ITU-T, Recommendation i.113. Vocabulary of Terms for Broadband Aspects of ISDN, Vol. Rev. 1, Geneva, 1991.
- 3. ATM Forum, User-Network Interface (UNI) Specification Version 3.1, 1994.
- ATM Forum, Traffic Management Specification Version 4.0, 1996.
- D. D. Clark, S. Shenker, and L. Zhang, Supporting realtime applications in an integrated services packet network: Architecture and mechanism, *Proc. SIGCOMM* 92, 1992, pp. 14-26.
- T. Chen, Evolution to the programmable Internet, *IEEE Commun. Mag.* 38: 124–128 (2000).
- 7. G. Eichler, Implementing integrated and differentiated services for the Internet with ATM networks: A practical approach, *IEEE Commun. Mag.* **38**: 132–141 (2000).

2430 STATISTICAL MULTIPLEXING

- D. Awduche, MPLS and traffic engineering in IP networks, IEEE Commun. Mag. 37: 42-47 (1999).
- L. Zhang et al., RSVP: A new resource reservation protocol, IEEE Network 7(5): 8–18 (1993).
- V. S. Frost and B. Melamed, Traffic modeling for telecommunications networks, *IEEE Commun. Mag.* 32(4): 70-81 (1994).
- D. Jagerman, B. Melamed, and W. Willinger, Stochastic modeling of traffic processes, in J. Dshalalow, ed., Frontiers in Queuing: Models, Methods and Problems, CRC Press, 1996.
- A. K. Erlang, The theory of probabilities and telephone conversations, Nyt Tidsskrift Matematik B 20: 33 (1909). (English translation in E. Brockmeyer, H. L. Halstrom and A. Jensen (1948), The life and works of A. K. Erlang, The Copenhagen Telephone Company, Copenhagen.
- E. G. Coffman and R. C. Wood, Interarrival statistics for time sharing systems, *Commun. ACM* 9: 5000-5003 (1966).
- P. E. Jackson and C. D. Stubbs, A study of multi-access computer communications, AFIPS Conf. Proc. 34: 491–504 (1969).
- E. Fuchs and P. E. Jackson, Estimates of distributions of random variables for certain computer communications traffic models, *Commun. ACM* 13: 752-757 (1970).
- P. F. Pawlita, Traffic measurements in data networks, recent measurement results, and some implications, *IEEE Trans. Commun.* 29(4): 525-535 (1981).
- R. Jain and S. A. Routhier, Packet trains- measurements and a new model for computer network traffic, *IEEE J. Select. Areas Commun.* 4(6): 986–995 (1986).
- J. F. Schoch and J. A. Hupp, Performance of the Ethernet local network, *Commun. ACM* 23(12): 711–721 (1980).
- D. N. Murray and P. H. Enslow, An experimental study of the performance of a local area network, *IEEE Commun. Mag.* 22: 48-53 (1984).
- F. A. Tobagi, Modeling and measurement techniques in packet communication networks, *Proc. IEEE* 66: 1423–1447 (1978).
- V. Paxson and S. Floyd, Wide-area traffic: The failure of Poisson modeling, *IEEE/ACM Trans. Network.* 3(3): 226-244 (1996).
- R. Caceres, P. Danzig, S. Jamin, and D. Mitzel, Characteristics of wide-area Tcp/Ip conversations, *Proc. ACM/SIG-COMM*, 1991, pp. 101–112.
- 23. K. S. Meier-Hellstern, P. E. Wirth, Y. Yan, and D. A. Hoeflin, Traffic models for ISDN data users: Office automation application, in A. Jensen and V. B. Iversen, eds., *Teletraffic and Data Traffic, a Period of Change*, Elsevier Science Publishers, 1991, pp. 167–172.
- R. Gusella, Characterizing the variability of arrival processes with indexes of dispersion, *IEEE J. Select. Areas Commun.* 9(2): 203-211 (1991).
- W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Network.* 2(1): 1-15 (1994).
- A. Erramilli, R. P. Singh, and P. Pruthi, Chaotic maps as models of packet traffic, *Proc. 14th Int. Teletraffic Congress*, 1994, Vol. 1, pp. 329–338.
- 27. W. Willinger, M. S. Taqqu, and A. Erramilli, A bibliographical guide to self-similar traffic and performance modeling

for modern high-speed networks, in F. P. Kelly, S. Zachary, and I. Ziedins, eds., *Stochastic Networks: Theory and Applications*, Oxford Univ. Press, 1996, pp. 339–366.

- R. B. Cooper, Introduction to Queueing Theory, 3rd ed., CEEPress Books, 1990.
- K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, *IEEE J. Select. Areas Commun.* 4(6): 833–846 (1986).
- H. Heffes and D. M. Lucantoni, A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE J. Select. Areas Commun.* 4: 856-868 (1986).
- A. E. Eckberg, Jr., Generalized peakedness of teletraffic processes, 10th Int. Teletraffic Congress, 1983.
- 32. A. E. Eckberg, Jr., Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness, in 11th Int. Teletraffic Congress, 1985.
- D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Chapman & Hall, 1966.
- A. A. Fredericks, Congestion in blocking systems a simple approximation technique, *Bell Syst. Tech. J.* 59: 805–827 (1980).
- M. Livny, B. Melamed, and A. K. Tsiolis, The impact of autocorrelation on queueing systems, *Manage. Sci.* 39(3): 322–339 (1993).
- W. Roberts, ed., *Performance Evaluation and Design of Multiservice Networks*, COST 224 Final Report, Commission of the European Communities, 1992.
- 37. I. Norros, J. W. Roberts, A. Simonian, and J. T. Virtamo, The superposition of variable bit rate sources in an ATM multiplexer, *IEEE J. Select. Areas Commun.* 9(3): 378-387 (1991).
- J. W. Roberts and J. T. Virtamo, The superposition of periodic cell arrival streams in an ATM multiplexer, *IEEE Trans. Commun.* 39: 298-303 (1991).
- D. Anick, D. Mitra, and M. M. Sondhi, Stochastic theory of a data-handling system with multiple sources, *Bell Syst. Tech.* J. 8: 1871–1894 (1982).
- E. Cinlar, Introduction to Stochastic Processes, Prentice-Hall, Englewood Cliffs, NJ, 1975.
- H. Heffes, A class of data traffic processes-covariance function characterization and related queueing results, *Bell* Syst. Tech. J. 59: 897-929 (1980).
- W. Fischer and K. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook, *Perform. Eval.* 18: 149-171 (1992).
- S. Q. Li and C. L. Hwang, Queue response to input correlation functions: Discrete spectral analysis, *IEEE/ACM Trans. Network.* 1(5): 522-533 (1993).
- 44. S. Q. Li and C. L. Hwang, Queue response to input correlation functions: Continuous spectral analysis, *IEEE/ACM Trans. Network.* 1(6): 678–692 (1993).
- 45. M. Decina, W. S. Gifford, R. Potter, and A. A. Robrock (guest eds.), Special issue on Integrated Services Digital Network: Recommendations and Field Trials—I, *IEEE J. Select. Areas Commun.* 4(3): (1986).
- 46. J. G. Gruber and N. H. Le, Performance requirements for integrated voice/data networks, *IEEE J. Select. Areas Commun.* 1(6): 981-1005 (1983).

- J. G. Gruber, Delay related issues in integrated voice and data networks, *IEEE Trans. Commun.* 29(6): 786–800 (1980).
- 48. N. Janakiraman, B. Pagurek, and J. E. Neilson, Performance analysis of an integrated switch with fixed or variable frame rate and movable voice/data boundary, *IEEE Trans. Commun.* **32**: 34–39 (1984).
- A. G. Konheim and R. L. Pickholtz, Analysis of integrated voice/data multiplexing, *IEEE Trans. Commun.* 32(2): 140-147 (1984).
- K. Sriram, P. K. Varshney, and J. G. Shantikumar, Discrete-time analysis of integrated voice-data multiplexers with and without speech activity detectors, *IEEE J. Select. Areas Commun.* 1(6): 1124-1132 (1983).
- H. H. Lee and C. K. Un, Performance analysis of statistical voice/data multiplexing systems with voice storage, *IEEE Trans. Commun.* 33(8): 809-819 (1985).
- T. Bially, B. Gold, and S. Seneff, A technique for adaptive voice flow control in integrated packet networks, *IEEE Trans. Commun.* 28: 325-333 (1980).
- E. A. Harrington, Voice/data integration using circuit switched networks, *IEEE Trans. Commun.* 28: 781-793 (1980).
- 54. C. J. Weinstein, M. K. Malpass, and M. J. Fisher, Data traffic performance of an integrated circuit and packetswitched multiplex structure, *IEEE Trans. Commun.* 28: 873-877 (1980).
- 55. B. Maglaris and M. Schwartz, Performance evaluation of a variable frame multiplexer for integrated switched networks, *IEEE Trans. Commun.* 29(6): 801-807 (1981).
- C. Roberge and J. Adoul, Fast on-line speech/voicebanddata discrimination for statistical multiplexing of data with telephone conversations, *IEEE Trans. Commun.* 34(8): 744-751 (1986).
- 57. P. T. Brady, A statistical analysis of on-off patterns in 16 conversations, *Bell Syst. Tech. J.* **47**: 73-91 (1968).
- P. T. Brady, A model for generating on-off speech patterns in two-way conversations, *Bell Syst. Tech. J.* 48: 2445–2472 (1969).
- K. Sriram and D. M. Lucantoni, Traffic smoothing effects of bit dropping in a packet voice multiplexer, *IEEE Trans. Commun.* 37(7): 703-712 (1989).
- B. G. Haskell, Buffer and channel sharing by several interframe picturephone coders, *Bell Syst. Tech. J.* 51(1): 261–289 (1972).
- M. Ghanbari, Two-layer coding of video signals for VBR networks, *IEEE J. Select. Areas Commun.* 7: 771–781 (1989).
- S. Tubaro, A two layers video coding scheme for ATM networks, Signal Process. Image Commun. 3: 129–141 (1991).
- R. Aravind, M. R. Civanlar, and A. R. Reibman, Packet loss resilience of mpeg-2 scalable coding algorithms, *IEEE Trans. Circuits Syst. Video Technol.* 6: 426–435 (1996).
- B. Maglaris et al., Performance models of statistical multiplexing in packet video communications, *IEEE Trans. Commun.* 36(7): 834-844 (1988).
- P. Sen, B. Maglaris, N. Rikli, and D. Anastassiou, Models for packet switching of variable bit-rate video sources, *IEEE J. Select. Areas Commun.* 7(5): 865–869 (1989).

- 66. R. Grunenfelder, J. P. Cosmos, S. Manthorpe, and A. Odinma-Okafor, Characterization of video codecs as autore-gressive moving average processes and related queueing system performance, *IEEE J. Select. Areas Commun.* 9: 284–293 (1991).
- 67. F. Yegenoglu, B. Jabbari, and Y. Zhang, Motion classified autoregressive modeling of variable bit rate video, *IEEE Trans. Circuits Syst. Video Technol.* 3: 42–53 (1993).
- D. Heyman, A. Tabatbai, and T. V. Lakshman, Statistical analysis and simulation study of video teletraffic in atm networks, *IEEE Trans. Circuits Syst. Video Technol.* 2: 49-59 (1992).
- D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, Methods for performance evaluation of VBR video traffic models, *IEEE/ACM Trans. Network.* 2: 176–180 (1994).
- D. P. Heyman and T. V. Lakshman, Source models for VBR broadcast-video traffic, *IEEE/ACM Trans. Network.* 4: 40-48 (1996).
- P. Skelly, M. Schwartz, and S. Dixit, A histogram based model for video traffic behavior in an ATM multiplexer, *IEEE/ACM Trans. Network.* 1(4): 447-459 (1993).
- P. Pancha and M. El Zarki, Prioritized transmission of VBR MPEG video, Proc. GLOBECOM'92, 1992, pp. 1135-1139.
- 73. M. R. Ismail, I. E. Lambadaris, M. Devetsikiotis, and A. R. Raye, Modelling prioritized MPEG video using tes and a frame spreading strategy for transmission in ATM networks, *Proc. INFOCOM'9*, 1995, Vol. 5, pp. 762–769.
- A. R. Reibman and A. W. Berger, Traffic descriptors for VBR video teleconferencing, *IEEE/ACM Trans. Network.* 3: 329-339 (1995).
- H. Heeke, A traffic control algorithm for ATM networks, *IEEE Trans. Circuits Syst. Video Technol.* 3(3): 183–189 (1993).
- P. Pancha and M. El. Zarki, Bandwidth allocation schemes for variable bit rate MPEG sources in ATM networks, *IEEE Trans. Circuits Syst. Video Technol.* 3(3): 192–198 (1993).
- 77. M. R. Frater, J. F. Arnold, and P. Tan, A new statistical model for traffic generated by VBR coders for television on the broadband isdn, *IEEE Trans. Circuits Syst. Video Technol.* 4(6): 521–526 (1994).
- M. Krunz and H. Hughes, A traffic model for MPEGcoded VBR streams, *Perform. Eval. Rev. (Proc. ACM* SIGMETRICS'95) 23: 47–55 (1995).
- K. Chandra and A. R. Reibman, Modeling one- and two-layer variable bit rate video, *IEEE/ACM Trans. Networks* 7(3): 398-413 (1999).
- M. W. Garrett and W. Willinger, Analysis, modeling and generation of self-similar VBR video traffic, *Proc. ACM* SIGCOMM'94, 1994, pp. 269-280.
- J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, Long range dependence in variable bit-rate video traffic, *IEEE Trans. Commun.* 43: 1566–1579 (1995).
- B. K. Ryu and A. Elwalid, The importance of long-range dependence of VBR video traffic in atm traffic engineering: Myths and realities, *Proc. ACM SIGCOMM'96*, 1996, pp. 3-14.
- M. Grossglauser and J. D. Bolot, On the relevance of long-range dependence in network traffic, *Proc. ACM* SIGCOMM'96, 1996, pp. 15-24.

- D. Heyman and T. V. Lakshman, What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Trans. Networking* 4: 301-317 (1996).
- K. Chandra and A. R. Reibman, Modeling two-layer SNR scalable MPEG-2 video traffic, Proc. 7th Int. Workshop Packet Video, 1996, pp. 7-12.
- G. Hasslinger, Semi-markovian modelling and performance analysis of variable rate traffic in ATM networks, *Telecommun. Syst.* 7: 281–298 (1997).
- L. Kleinrock, *Queueing Systems*, Vol. 2, Wiley, New York, 1976.
- J. Walrand and P. Varaiya, *High-Performance Communica*tion Networks, Morgan Kaufmann, 1996.
- 89. M. Schwartz, Broadband Integrated Networks, Prentice-Hall, 1996.
- 90. J. W. Mark and S.-Q. Li, Traffic characterization for integrated services networks, *IEEE Trans. Commun.* 38: 1231-1242 (1990).
- C. Thompson, K. Chandra, S. Mulpur, and J. Davis, Packet delay in multiplexed video streams, *Telecommun. Syst.* 16: 335-345 (2001).
- 92. J. Abate, G. L. Choudhury, and W. Whitt, Asymptotics for steady-state tail probabilities in structured Markov queueing models, *Stochastic Models* 10: 99-143 (1994).
- R. G. Addie and M. Zukerman, An approximation for performance evaluation of stationary single server queues, *IEEE Trans. Commun.* 42: 3150-3160 (1994).
- 94. A. Elwalid et al., Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing, *IEEE J. Select. Areas Commun.* 13: 1004–1016 (1995).
- 95. J. Choe and N. B. Shroff, A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks, *IEEE/ACM Trans. Network.* 6(5): 659-671 (1998).
- 96. I. Norros, On the use of Fractal Brownian motion in the theory of connectionless networks, *IEEE J. Select. Areas Commun.* 13: 953–962 (1995).
- A. I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, *IEEE / ACM Trans. Network.* 1(3): 329-343 (1993).
- G. Choudhury, D. M. Lucantoni, and W. Whitt, Squeezing the most of ATM, *IEEE Trans. Commun.* 44(2): 203-217 (1996).
- E. W. Knightly and N. B. Schroff, Admission control for statistical qos: Theory and practice, *IEEE Network* 13(2): 20-29 (1999).
- 100. D. Ferrari and D. Verma, A scheme for real-time channel establishment in wide-area networks, *IEEE J. Select. Areas Commun.* 8: 368–379 (1990).
- 101. G. Kesidis, J. Walrand, and C. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, *IEEE/ACM Trans. Network.* 1(4): 424–428 (1993).
- 102. C. S. Chang and J. A. Thomas, Effective bandwidth in highspeed digital networks, *IEEE J. Select. Areas Commun.* 13: 1019–1114 (1995).
- F. Kelly, Stochastic Networks: Theory and Applications, Oxford Univ. Press, 1996.

- 104. C. S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, *IEEE Trans. Automatic Control* 39: 913-931 (1994).
- 105. V. G. Kulkarni, L. Gun, and P. F. Chimento, Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer, *IEEE J. Select. Areas Commun.* 6(13): 1039–1047 (1995).

STREAMING VIDEO

ROBERT A. COHEN Rensselaer Polytechnic Institute Troy, New York

HAYDER RADHA Michigan State University East Lansing, Michigan

1. INTRODUCTION

In the early 1990s, as the Internet and the World Wide Web were becoming ubiquitous in the office and home, users would download compressed digital video files onto their computers. Once downloaded, these clips would be viewed by opening the video files with a compatible player program on the computer. These video files often required several minutes or hours to download, especially through slow modem connections over telephone lines. Depending on which compression scheme was used, an incomplete video file may not have been viewable at all. If a connection were not reliable, the download process would have to be repeated, which forced the user to wait even longer to view the clip. This dilemma was solved through the use of videostreaming.

Videostreaming is the transmission, at a regulated rate, of digital video over a network from a server to a client (player) in such a way that the client can display the video while the video is still being transmitted. In other words, the client can start playing the video without having to wait for the entire clip to be downloaded. Using an assortment of compression techniques, packetization, and transport protocols, video can be manipulated to be streamed over a wide variety of networks. Publicly available methods for streaming video over the Internet first came about in 1994 using the Multicast Backbone (MBone) [1], in which many users could simultaneously receive multimedia in real time. Several commercial videostreaming solutions first became available between 1995 and 1997. Between 1997 and today, these streaming systems have evolved to address a variety of streaming needs, ranging from small personal streaming applications all the way to large distributed streaming systems designed for thousands of viewers. Several articles describing the many technologies and current trends in videostreaming can be found in Civanlar et al. [2]. Information regarding the histories and uses of specific commercial streaming systems can also be found [3-5].

The purpose of this article is to give an overview of how videostreaming works, the various technical aspects of videostreaming, video and network-related issues involved in streaming, technical solutions to these issues, and future directions of this field.

2. VIDEOSTREAMING SYSTEM DESCRIPTION AND OPERATION

The two main components of a videostreaming system are the streaming server and the client. The *server*, which can range in size from a small PC-like system with a camera to a large collection of networked computers, streams video into the network. The *client* is the device or computer on which the video is displayed. In bidirectional systems such as those used with videoconferencing, the server and client are combined into one device. The focus of this chapter will be on one-way videostreaming. For more information on videoconferencing, see Schaporst [6]. For an in-depth look at streaming video over the Internet, see [7].

2.1. Video Server Architecture

The components of a typical videostreaming server are shown in Fig. 1. Digitized video typically has a very high bit rate. For example, uncompressed 320×240 -pixel RGB (red-green-blue) video at 30 frames per second (FPS) requires a transmission bandwidth of over 55 Mbps, which is well over the bandwidth available on most of the networks in use today. A video encoder therefore is used to compress the program into a lower-rate bitstream. Details on video coders that are used for streaming are given in Section 3. Once the video is compressed, it can be streamed live over the network, or it can be stored for streaming later on demand.

The session controller chooses a program to be streamed. It typically receives commands from the client, which is shown in Fig. 2. Issues related to session control and management are discussed in Section 5. Once a program is selected, the rate of the compressed video data will be chosen on the basis of data from the session controller and from the network and client limitations. On the basis of the performance of the network, the congestion controller can also control the rate at which a video program is being streamed. Since video is usually streamed over packet-based networks, it must be packetized in a way that is appropriate given the type of network and video encoding method. Section 4 gives more information on packetization schemes and networkrelated issues. If the video is going to be streamed over a lossy network, error control data can be added to the video packets to reduce the effects of packet loss or delay. These application-layer packets are then fed to the transport layer, which uses protocols such as UDP or ATM to send data over the network.

2.2. Video Client Architecture

A client system is shown in Fig. 2. The user typically first chooses a program to be viewed. The program list is usually sent over the network via a separate application, such as a Web browser. Once the program is chosen, the client's session controller tells the server's controller to start streaming the video. The client receives and demultiplexes the packets to reconstruct the encoded videostream. If errors are detected, the client will, if possible, correct the errors before sending the reconstructed stream to the video decoder. As described later, error detection information can be fed back to the server for congestion and rate control. The video decoder then decompresses the stream to produce video that can be displayed. In some cases, the decoder can try to conceal uncorrectable errors so the user does not see too many distracting artifacts in the displayed video.

3. VIDEO CODING FOR STREAMING

3.1. Basics of Video Coding

As discussed above, video coding is needed to reduce the amount of information transmitted over a streaming session. The basic objective of any video coding method is to reduce or virtually eliminate the amount of redundant data contained within a series of pictures. Usually, each picture contains a large number of pixels that are very similar. This is particularly true for pixels located within a small neighborhood of the picture. Therefore, the total number of pixels within a picture can be represented by a smaller number of bytes. Moreover, consecutive pictures within a video sequence are very similar. Consequently,



Figure 1. Videostreaming server.



Figure 2. Videostreaming client.

there are two types of redundancies within a video sequence that a video encoder exploits to reduce the amount of transmitted data: (1) redundancies within each picture and (2) redundancies among consecutive pictures. Redundancies among consecutive pictures are known as *temporal redundancies* or *interframe* (or *interpicture*) *redundancies*. When coding a new picture, temporal redundancies are reduced through prediction from one or more previously coded pictures. The most popular picture prediction approach is based on motion estimation/motion compensation (ME/MC) using block matching (BM) among adjacent pictures [8].

International video coding standards such as those from MPEG [9–11], H.261 [12], and H.263 [13] employ *intracoded pictures* (I frames), which do not depend on prediction from other pictures. Any video sequence requires a minimum of one I frame, which is usually the first picture of the sequence. The abovementioned standards also use *prediction frames* (P frames), which depend on a previously coded I or P frame. This results in the picture coding structure shown in Fig. 3. MPEG-2 [10], MPEG-4 [11], and other more recent compression methods employ *bidirectional prediction frames* (B frames), which use prediction from a previously coded I or P frame *and* a future P frame as shown in Fig. 3.

The pixels of an I frame or the *residual pixels* of P and B frames are usually coded using transform-domain methods. The *discrete cosine transform* (DCT) is the most popular transform coding method used for video compression [14]. MPEG-1 [9], MPEG-2, H.261, and H.263

are all based on a DCT coding method. Another popular transform that has received a great deal of attention is the wavelet transform [15]. For example, JPEG-2000 [16] is based on a wavelet transform coding method.

3.2. Scalable Video Coding for Streaming Applications

Scalable video coding is a desirable tool for many multimedia applications and services. Video scalability, for example, can be used in systems employing decoders with a wide range of processing power. In this case, processors with low computational power decode only a subset of the scalable videostream. Another use of scalable video is in environments with a variable and unpredictable transmission bandwidth (e.g., the Internet or wireless networks). In this case, receivers with low access bandwidth receive, and consequently decode, only a subset of the scalable videostream, where the amount of that subset is proportional to the available bandwidth. Without scalability, a videostream could be rendered useless if the network bandwidth drops below the coding rate.

Several video scalability approaches have been adopted by leading video compression standards such as MPEG-2, MPEG-4, and H.263. Temporal, spatial, and quality [signal-to-noise ratio (SNR)] scalability types have been defined in these standards. Temporal scalability applies to the frame rate (in frames per second) of the video. With spatial scalability, the size or resolution of each video frame can vary. In SNR-scalable systems, the quality of the video can be increased or decreased. All of these standardized types of scalable video consist of a *base layer*



Figure 3. Examples of prediction structures used for video coding.

(BL) and one or multiple *enhancement layers* (ELs). The BL part of the scalable videostream represents, in general, the minimum amount of data needed for decoding a viewable video sequence from the stream. The EL part of the stream represents additional information, and therefore enhances the video signal representation when decoded by the receiver.

For each type of video scalability, a certain *scalability structure* is used. The scalability structure defines the relationship among the pictures of the BL and the pictures of the enhancement layer. Figure 4 illustrates examples of video scalability structures. MPEG-4 also supports object-based scalability structures for arbitrarily shaped video objects.

Another type of scalability, which has been primarily used for coding still images, is *fine-granular scalability* [17]. Images coded with this type of scalability can be decoded progressively. In other words, the decoder can start decoding and displaying the image after receiving a very small amount of data. As more data are received, the quality of the decoded image is progressively enhanced until the complete information is received, decoded, and displayed. Among leading international standards, progressive image coding is one of the modes supported in JPEG-2000 and the still-image coding tool in MPEG-4 video.

When compared with nonscalable methods, a disadvantage of scalable video compression is coding efficiency. In order to increase coding efficiency, video scalability methods normally rely on relatively complex structures (such as the spatial and temporal scalability examples shown in Fig. 4). By using information from as many pictures as possible from both the BL and EL, coding efficiency can be improved when compressing an enhancement-layer picture. However, using prediction among pictures within the enhancement layer either eliminates or significantly reduces the fine-granular scalability feature, which is desirable for environments with a wide range of available bandwidth (e.g., the Internet).

3.2.1. MPEG-4 Fine-Granular Scalability (FGS) Video Coding. In order to strike a balance between coding efficiency and fine-granularity requirements, a more recent activity in MPEG-4 adopted a hybrid scalability structure characterized by a DCT motion-compensated base layer and a fine-granular scalable enhancement layer. This scalability structure is illustrated in Fig. 5.

The base layer carries a minimally acceptable quality of video to be reliably delivered using a, packet-loss recovery method such as retransmission. The enhancement layer improves the base layer video by fully utilizing the bandwidth available to individual clients. By employing a motion-compensated base layer, coding efficiency from temporal redundancy exploitation is partially retained. The base and a single-enhancement layer streams can be either stored for later transmission, or can be directly streamed by the server in real time. The encoder generates a compressed bitstream that can be transmitted over any bit rate available over the range of bandwidth $[R_{\min}, R_{\max}]$. The base-layer bit rate has to meet the following constraint: $R_{\rm BL} \leq R_{\rm min}$. The enhancement layer is overcoded using a bit rate $(R_{\text{max}} - R_{\text{BL}})$. It is important to note that the range $[R_{\min}, R_{\max}]$ can be determined offline for a particular set of Internet access technologies.



Temporal scalability

Figure 4. Examples of video scalability structures.



DCT base layer

Figure 5. Video scalability structure with fine granularity.

For example, $R_{\rm min} = 20$ kbps and $R_{\rm max} = 100$ kbps can be used for analog modem/ISDN access technologies. More sophisticated techniques can also be employed in real-time to estimate the range $[R_{\rm min}, R_{\rm max}]$. For unicast streaming, the system estimates, in real time, the available bandwidth R for a particular session. On the basis of this estimate, the server transmits the enhancement layer using a bit rate $R_{\rm EL}$:

$$R_{\rm EL} = \min(R_{\rm max} - R_{\rm BL}, R - R_{\rm BL})$$

Because of the fine granularity of the enhancement layer, its real-time rate control aspect can be implemented with minimal processing. For multicast streaming, a set of intermediate bit rates R_1, R_2, \ldots, R_N can be used to partition the enhancement layer into substreams. In this case, N fine-granular streams are multicasted using the bit rates:

$$R_{e1} = R_1 - R_{\rm BL}, R_{e2} = R_2 - R_1, \dots, R_{eN} = R_N - R_{N-1}$$

where $R_{
m BL} < R_1 < R_2 \dots < R_{N-1} < R_N \le R_{
m max}.$

One can choose from many alternative compression methods when coding the BL and EL layers of the FGS structure shown in Fig. 5. For the FGS MPEG-4 standard, the base layer is coded using a DCT-based set of video compression tools. The FGS MPEG-4 enhancement layer is coded using an embedded DCT coding scheme. MPEG-4 FGS also support temporal scalability and hybrid SNR/temporal scalabilities. For more details regarding the MPEG-4 FGS scalable video method and many of its coding tools, the reader is referred to the paper by Radha et al. [18]. FGS is also very resilient to packet losses [19], which are common in streaming applications over the Internet.

4. PACKETIZATION AND TRANSPORT-LAYER ISSUES

Once coded, the next step in streaming is to transmit the compressed video over a network. Underlying network transport protocols such as TCP, UDP, or ATM [20] work well for transferring data over networks, but they are not very effective alone as the only packetization and fragmentation methods for streaming time-dependent media such as video or audio. When ignored, factors such as end-to-end delay, packet loss, delay jitter, network bandwidth bottlenecks, congestion, buffering, and decoder complexity/capability all can render a videostream useless.

4.1. Application-Layer Packetization

Many of the abovementioned issues can be addressed by intelligently packetizing video data at the application layer prior to sending it through the network transport layer using a transport protocol. Forming packets by breaking the data at their natural separation points is known as application-layer framing [21]. Frame, slice [10], or DCT block boundaries are good natural separation points for packetizing videostreams. In most cases, if properly framed packets are lost during streaming, the client will still be able to make use of the other packets to decode a viewable program. These application-layer packets can be created and stored prior to streaming, or they may be generated in real time as the video is being streamed.

At the application layer, framing alone, however, will not solve timing and synchronization issues in videostreaming. Transport protocols such as TCP and UDP can determine sequential relations between packets, but they have nothing to resolve real-time temporal relationships. Adding time indicators, or *timestamps*, to application-layer packets allows the client to know the temporal relationships between video packets. Timestamps allow clients to do things such as synchronize media streams, throw out packets that arrive too late to be decoded and displayed, and manage buffer control issues. One popular method of application-layer framing that has the above-mentioned features is the Real-Time Transport Protocol (RTP). Another standard that addresses properties of multimedia streams is Quicktime [22].

4.2. The Real-Time Transport Protocol (RTP)

The Real-Time Transport Protocol (RTP) [23] is a packetization protocol that is most commonly used at the application layer to packetize time-dependent data such as video or audio. RTP packetization is often done prior to transport-layer packetization such as UDP, so that RTP handles media-dependent issues such as timing. synchronization, and multiplexing, and UDP handles network-dependent issues such as data packet framing and multicasting. An example of this kind of multilevel packetization is shown in Fig. 6. Like other network protocols, RTP consists of a header and a payload. The RTP header contains bit fields to represent information such as sequencing, source identification, and timing. Detailed descriptions of the RTP header may be found elsewhere [23,24]. A few of the fields that are of particular relevance to videostreaming are:

Timestamp. This 32-bit field represents a sampling of a clock consistent with the type of data being


Figure 6. Multilayer packetization.

streamed. For frame-based video, every RTP packet containing data from the same video frame will have the same timestamp. For MPEG video, the timestamp typically has a resolution of 90 kHz.

- *Marker bit.* A complete frame of compressed video may still be too large to put into one packet. If a video frame is split into multiple RTP packets, the marker bit typically is set for the last packet in the video frame. By setting the marker bit in the last packet, the client does not have to wait for the packet containing the start of the next frame to tell us that we can decode the current frame.
- Sequence number. Since all packets of a single video frame have the same RTP timestamp, the 16-bit RTP sequence number, which is incremented for each RTP packet sent, can be used to determine the proper order of packets in a video frame. This way, the client does not have to look at the video data in the payload to determine the proper RTP packet sequence.

4.3. Packet Size Considerations

If the length of a packet set through the transport layer is larger than the network's maximum transmission unit (MTU) [20], the packet will be fragmented. It is therefore desirable to ensure that the RTP packet size is less than the network MTU so that the RTP packet will not be split into multiple packets. In the example shown in Fig. 6, the first video frame is too large to fit into one MTU, so it is split into two RTP packets. The obvious splitting method would be to simply fill one packet so that the resulting network packet size is just below the MTU value, and then put whatever is left over into the next packet. Since a transmitted packet may be delayed or lost, however, it is better to use application-layer framing, and choose a splitting point (or points) that take the video structure into account. For example, with MPEG-based video, the split could be done at a slice boundary. If, for example, the second packet of a two-packet frame is lost, the client will not have to throw away data in the first packet since they can be decoded into a viewable partial picture (assuming that a partial picture is something that the client would want to display). As another example, for multilayer scalable video, large frames could be split at layer boundaries so that a missing packet will have a lower probability of making the received packets for a frame unusable.

Large variations in packet size can also cause problems for videostreaming. In MPEG-2 and MPEG-4, I frames are typically large, while P and B frames are relatively small. Suppose that a 10-FPS video program is being streamed. The average time needed to transmit one frame would therefore be 0.1 s. The I frame, however, could require more than 0.1 s to stream, and the P and B frames would use less than 0.1 s. If, for example, the Group of Pictures (GOP) size (i.e., I frame period) is 2 s, we would have one I frame and 19 P or B frames in every GOP. If this video is coded so that the I frame is very large, the server could take over one second to transmit the I frame. This amount of time has the potential to cause problems if the decode buffer in the client is close to being empty, since a large I frame won't be available for display until all the packets in the frame have been received. It is therefore prudent for the content creator to use an encoding rate control method that is amenable toward streaming. The rate control method that results in the best compression ratio may not necessarily be the best method to use for streaming.

5. END-TO-END SESSION CONTROL AND MANAGEMENT

As described earlier, continuous decoding and playback of video at the receiver characterizes videostreaming and differentiates it from traditional applications such as email, FTP, and Webpage download. Consequently, additional challenges of videostreaming are to establish and maintain a regulated and continuous transmission of video data between the server and the client. This makes end-to-end session and rate control crucial aspects of any videostreaming solution.

5.1. Session Control

Some of the basic steps in multimedia session control are as follows:

- Choose a program to be streamed.
- Choose a rate at which the program will be streamed.
- Identify the network architecture over which the streaming will be done (e.g., single vs. multiple clients, unicast vs. multicast, UDP vs. http).

- Address other factors such as error control, media property rights, and encryption.
- Control the positioning, starting, stopping, and playing of the media stream.
- Adapt the streaming properties on the basis of various conditions related to the server, client, or network.

Some of the session control mechanisms used by today's commercially available streaming solutions are proprietary. An example of a publicly specified session control scheme is RTSP [25], which was designed for use over the Internet. A fundamental overview on how multimedia streams can be controlled over the Internet can be found in a paper by Schulzrinne [26].

The bidirectional nature of session control must also be taken into consideration when designing a streaming system. If the round-trip time of a network is not too large, proprietary methods or public protocols such as the RTP Control Protocol (RTCP) [23] can be used to help the server to adapt to varying network conditions and client configurations. In RTCP, sender and receiver reports are used to feedback timing, quality, and other related quantities to allow the server to adapt accordingly.

An aspect of session control that relates to video coding is the choice of rates used for streaming. If a viewer chooses to stream a nonscalable 128-kbps video program over a 56-kbps connection, a long startup delay may be incurred, as explained in the next section. One way commonly used to solve this problem is to encode the video at different rates, and store the multiple streams in different files. This method would allow the client (either automatically or user-driven) to choose the stream best suited for the given network connection. Encoding a video program at several different rates in separate files can be time-consuming, and it can create disk-space and file management problems for the content creator. To improve the server/client system, switched-rate streaming can be used. In switched-rate streaming, the video is encoded into a file that can be streamed at different rates. This way, the content creator can generate a file that can be streamed, for example, at 28, 56, and 128 kbps (and perhaps a few rates in between), and then the server can switch rates during streaming on the basis of feedback from the client. The problem of choosing rates during streaming can also be solved by using scalable video, as described in Section 3. If the video is continuously scalable, as in FGS, a variety of methods can be used to choose the appropriate packet size and streaming rate for the given connection [27].

5.2. Playout Buffer and Delay Considerations

If video were being transmitted or streamed over a guaranteed constant-rate network, the buffer internal to the video decoder (e.g., the decoder buffer in MPEG) would be sufficient to ensure a consistent viewing experience for the end user. Networks such as the Internet, however, are affected by loss and delays, which necessitate the use of a larger buffer at the client to store enough video so that the program will continue to be played on the client during these periods of rate variation, delay, or packet loss. A playout buffer is therefore used to store the videostream as it is being received. This buffer isolates the decoder from the network.

There is a tradeoff between the size of the playout buffer and the amount of delay incurred from when the server starts streaming to when the video is actually displayed on the client. This delay is also encountered when the playout buffer empties due to network congestion, and therefore must be refilled with an appropriate amount of video. One extreme way to abate network effects is to make the playout buffer large enough to contain the entire videostream, and then wait for this buffer to fill with the entire program before playing. This solution, of course, is not acceptable since the whole idea of streaming is to allow us to view the program without waiting for all of it to be sent to the client. If we initially fill the playout buffer with only a second or two of video, the viewer will be less likely to be annoyed by long waits. Having a short playout delay, however, increases the likelihood of having the buffer empty, resulting in more frequent occurrences of video pausing due to buffer refills.

If video is being streamed in real time (i.e., at the rate of the encoded video), then it would take, for example, 10 s to fill the buffer with 10 s of video. If the streaming session is not rate-limited, and if the network allows, the initial buffer-filling data can be sent faster than video rates. This way, the viewer may have to wait only a few seconds to have 10 s of video in the buffer. This solution is used by some of today's commercially available streaming systems to reduce waiting times at the client. It is important to note that if this high-speed buffer filling is done too frequently, it greatly increases the data rate used over the network, so a server should be designed not to overwhelm the network in cases when the client asks for buffer refills too frequently.

5.3. Congestion Control

In particular, a streaming video session needs to estimate the effective available bandwidth between the server and the client. This bandwidth estimate can be used to regulate the rate at which video is transmitted over the end-to-end session. Bandwidth estimation over a shared network, such as the Internet, is a very challenging problem. More importantly, bandwidth estimation is crucial for eliminating "gridlock" or congestion over the shared network. In other words, the large numbers of sessions (streaming or other applications) that use the Internet simultaneously need to employ some mechanism that provides a fair usage of the shared resources of the Internet. Without such a mechanism, the different sessions would be transmitting data at rates higher than the available bandwidth. This would lead to congestion over the shared network and eventually may lead to some form of gridlock. Consequently, this aspect of videostreaming, or available bandwidth estimation, is commonly known by the networking and Internet community as *congestion control*.

It is important to note that, prior to the emergence of streaming applications, congestion control represented one of the cornerstones of TCP. In fact, many experts attribute the continuous success of the Internet and its growth to the ability of the TCP/IP protocol stack to provide a robust and scalable congestion control algorithm. The scalability of a congestion control algorithm is its ability to support a large number of sessions while (1) maintaining a robust level of fairness among these sessions and (2) eliminating the possibility of congestion (or gridlock) over the shared network.

The TCP congestion control algorithm is based on the following simple strategy. The server increases its sending rate linearly until a packet-loss event occurs. This event is detected by the receiver and communicated back to the server. Once a packet loss is detected, the server reduces its sending rate exponentially. Therefore, the TCP congestion control can be expressed as a linear increase–exponential decrease congestion control. Another popular characterization for TCP congestion control is additive increase/multiplicative decrease (AIMD) [28]. The original work in TCP congestion control is attributed to Jacobson [29].

Since the emergence of streaming applications, one of the key concerns has been the impact of these applications on the congestion of the Internet. In particular, streaming applications are supported by the UDP protocol rather than by TCP. UDP, however, does not provide any standardized congestion control mechanism. Consequently, for streaming applications, congestion control is the responsibility of the application. Therefore, real-time streaming applications that do not support some form of congestion control may become misbehaving in the sense that they may monopolize the available shared bandwidth over the Internet, and hence, they may become unfair to the mainstream (wellbehaving) TCP applications. So far, this concern has been addressed by popular streaming solutions, such as those from RealNetworks [30] and Microsoft [31], through a relatively straightforward method. Multiple streams are generated at different bit rates and stored at the server. For example, streams for 56-kbps modems as well as for ISDN and higher rates (e.g., cable-modem or DSL rates) are compressed and made available for users. The application relies on the user to select the appropriate stream at the beginning of the streaming session. This simple approach is not sustainable for the long term, especially if we anticipate that streaming applications will become increasingly popular. Consequently, researchers have proposed several approaches for congestion control of media streaming, in general, and videostreaming in particular.

One of the most popular congestion control strategies that have been proposed and studied thoroughly for streaming applications is what is known as *TCP-friendly congestion control* [32]. The basic premise of this approach is that a streaming application follows a congestion control mechanism that is similar to the congestion control mechanism employed by TCP. This naturally makes streaming applications *fair* to the more popular TCP applications in terms of sharing the available bandwidth over Internet sessions. This fairness explains the reason behind the label "TCP-friendly."

6. FUTURE DIRECTIONS IN VIDEOSTREAMING

Videostreaming is expected to continue its growth toward a mainstream Web application. This growth has to be accompanied with successful efforts in addressing some of the key challenges in video coding and congestion control strategies. Regarding video coding, the key challenge is to provide new solutions that address the need for (1) scalability (i.e., to address the bandwidth variation, devices and network heterogeneity, and related QoS issues over the Internet); (2) new functionality (e.g., interactivity); and (3) providing high-quality video. New trends in video coding, such as 3D motion-compensated wavelet compression [33], could provide some answers to these challenges. Multiple description coding [34] is being looked at to stream video over channels with different characteristics. In addition to streaming frames of video, the object-based coding capabilities of MPEG-4 could be used to stream objects from various sources, which are composited at the client for new kinds of interactive streaming experiences. Moreover, proxy-based services that provide some form of *transcoding* of video may be a viable approach for addressing the need for scalable and high quality video. In particular, a new framework known as TranScaling has been proposed [35] to address the scalability and video quality issues of videostreaming over the wireless Internet. Under TranScaling, a scalable stream is mapped at a gateway server to one or more scalable streams with higher-quality video.

Regarding congestion control, further studies are needed in the area of TCP-friendly algorithms and related approaches. It is important to note that TCPfriendly congestion control represents a class of (or an umbrella of) mechanisms that are friendly to TCP sessions. This class of mechanisms includes, for example, the AIMD congestion control strategy mentioned above. Therefore, different types of TCP-friendly algorithms provide different levels of performance in terms of fairness and scalability. For example, it has been shown that AIMD TCP-friendly congestion control has many optimal and desirable attributes when compared with other TCP-friendly algorithms [36]. Moreover, other and more general congestion control frameworks have been proposed for streaming applications. This includes equation-based congestion control [37], binomial [38], and ideally scalable [36] congestion control algorithms.

Error resilience is also becoming increasingly important, especially when video is streamed over specialized channels. As the use of wireless networks increases, adding error detection, correction, and concealment to videostreams becomes just as important as the method used to compress the video. MPEG-4, for example, has error resilience built into the standard [39]. Finally, protecting digital property rights has become a topic of interest for streaming copyrighted material. Property rights management is being added to the latest versions of some of the more popular commercially available streaming systems described earlier in this article. Another way to protect the video is through the use of watermarking, in which a nonvisible (or barely visible) signal is added to the video [40]. This watermark would be present in subsequent copies of the video, even if it is transcoded or recompressed. Solutions to all these challenges, combined with the increasing speed of network connectivity for end users, could soon make videostreaming a popular alternative to standard broadcast television.

BIOGRAPHIES

Robert Cohen received a B.S., summa cum laude, and a M.S., both in computer and systems engineering from Rensselaer Polytechnic Institute in Troy, New York. In 1990, he joined Philips Research in Briarcliff Manor, New York as a senior member of the research staff. At Philips Research, he was a project leader or team member doing research, development, patenting, and publishing in areas including video coding and signal processing, rapid prototyping for VLSI video systems, the Grand Alliance HDTV decoder, statistical multiplexing for MPEG video encoders, scalable MPEG-4 video streaming, and next-generation video surveillance systems. He has recently returned to Rensselaer Polytechnic Institute to complete his Ph.D. in electrical engineering. His current research interests include video coding and transmission, multimedia streaming, and image and video processing algorithms and architectures.

Hayder Radha received his Ph.D. ('93) and Ph.M. ('91) degrees from Columbia University, his M.S. degree from Purdue University in 1986, and his B.S. with honors degree from Michigan State University in 1984, all in electrical engineering. He joined Bell Laboratories in 1986 as a Member of Technical Staff. He worked at Bell Labs between 1986 and 1996 in the areas of digital communications, signal and image processing, and broadband multimedia communications. In 1996, he joined Philips Research as a principal member of research staff, and worked in the areas of video communications, networking, and high definition television. He initiated an Internet video research program at Philips Research and led a team of researchers working on scalable video coding, networking, and streaming algorithms. In 2000, he joined Michigan State University as an associate professor in the Department of Electrical and Computer Engineering.

He served as a cochair and an editor of the ATM and LAN Video Coding Experts Group of the ITU-T between 1994 and 1996. His research interests include image and video coding, multimedia communications and networking, and the transmission of multimedia data over wireless and packet networks. He has 25 patents in these areas (granted and pending). Dr. Radha received the Bell Laboratories Distinguished Member of Technical Staff Award and Appointment in 1993 and the Research Fellow Appointment at Philips Research in 2000. He is also a senior member of the IEEE.

BIBLIOGRAPHY

- 1. H. Eriksson, MBONE: The multicast backbone, Commun. ACM 36: 68-77 (Jan. 1993).
- M. R. Civanlar et al., guest eds., *IEEE Trans. Circuits Syst. Video Technol.* (Special Issue on Streaming Video) 11(3) (March 2001).

- 3. H. P. Alesso, e-Video: Producing Internet Video as Broadband Technologies Converge, Addison-Wesley Professional, 2000.
- G. J. Conklin et al., Video coding for streaming media delivery on the Internet, *IEEE Trans. Circuits Syst. Video Technol.* 11(3): 269–281 (March 2001).
- 5. J. Alvear, Web Developer.com Guide to Streaming Multimedia, Wiley Computer Publishing, New York, 1998.
- 6. R. Schaporst, Videoconferencing and Videotelephony: Technology and Standards, 2nd ed., Artech House, Boston, 1999.
- D. Wu et al., Streaming video over the Internet: approaches and directions, *IEEE Trans. Circuits Syst. Video Technol.* 11(3): 282–299 (March 2001).
- 8. A. M. Tekalp, *Digital Video Processing*, Prentice-Hall, Upper Saddle River, NJ, 1995.
- D. J. LeGall, MPEG: A video compression standard for multimedia applications, *Commun. ACM* 34: 46-58 (1991).
- 10. J. L. Mitchell et al., eds., MPEG Video Compression Standard Kluwer, 1996.
- 11. A. Puri and T. Chen, eds., *Multimedia Systems, Standards, and Networks*, Marcel Dekker, New York, 2000.
- Video Codec for Audiovisual Services at 64 kBit/s, CCITT Recommendation H.261, 1990.
- Video Coding for Low Bitrate Communications, ITU-T Recommendation H.263, Nov. 1995.
- A. N. Netravali and B. G. Haskell, *Digital Pictures—Repre*sentation and Compression, 2nd ed., Plenum Press, New York, 1995.
- S. Mallat, A Wavelet Tour of Signal Processing, 2nd ed., Academic Press, San Diego, 1999.
- D. S. Taubman and M. W. Marcellin, JPEG 2000: Image Compression Fundamentals, Standards, and Practices, Kluwer, 2001.
- H. Radha et al., Scalable Internet video using MPEG-4, Signal Process. Image Commun. 15: 95-126 (Sept. 1999).
- H. Radha, M. van der Schaar, and Y. Chen, The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP, *IEEE Trans. Multimedia* 3(1): 53-68 (March 2001).
- M. van der Schaar and H. Radha, Unequal packet loss protection for fine-granular-scalability video, *IEEE Trans. Multimedia* 3(4): 381-394 (Dec. 2001).
- W. R. Stevens, UNIX Network Programming, Vol. 1, 2nd ed., Networking APIs: Sockets and XTI, Prentice-Hall, Engelwood Cliffs, NJ, 1998.
- D. D. Clark and D. L. Tennehouse, Architecture considerations for a new generation of protocols, *Proc. ACM SIGCOMM* '90: 201–208, Sept. 1990.
- E. Hoffert et al., QuickTime: An extensible standard for digital multimedia, Proc. IEEE Compcon Spring '92, 1992, pp. 15-20.
- H. Schulzrinne et al., *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, Internet Engineering Task Force, Jan. 1996.
- 24. M.-T. Sun and A. R. Reibman, eds., Compressed Video over Networks, Marcel Dekker, New York, 2001.
- H. Schulzrinne, A. Rao, and R. Lanphier, *Real Time Stream*ing Protocol (*RTSP*), RFC 2326, Internet Engineering Task Force, Apr. 1998.

- H. Schulzrinne, A comprehensive multimedia control architecture for the Internet, Proc. IEEE Workshop on Network and Operating System Support for Digital Audio and Video, 1997, pp. 65-76.
- R. Cohen and H. Radha, Streaming fine-grained scalable video over packet-based networks, *Proc. IEEE GlobeCom* '00 1: 288-292 (Nov.-Dec. 2000).
- D.-M. Chiu and R. Jain, Analysis of the increase and decrease algorithms for congestion avoidance in computer networks, *Comput. Networks ISDN Syst.* 17: 1–14 (1989).
- V. Jacobson, Congestion avoidance and control, ACM SIG-COMM Comput. Commun. Rev. 18(4): 314–329 (Aug. 1988).
- RealNetworks home page (online), http://www.realnetworks. com.
- 31. Microsoft Windows Media Technologies Player home page (online), http://www.microsoft.com/windows/windows media.
- M. Allman, V. Paxson, and W. Stevens, *TCP Congestion Control*, RFC 2581, Internet Engineering Task Force, April 1999.
- S.-J. Choi and J. W. Woods, Motion-compensated 3-D subband coding of video, *IEEE Trans. Circuits Syst. Video Technol.* 8(2): 155-167 (Feb. 1999).
- J. K. Wolf, A. D. Wyner, and J. Ziv, Source coding for multiple descriptions, *Bell Syst. Tech. J.* 59(10): 1909–1921 (Dec. 1980).
- 35. H. Radha, TranScaling: A video coding and multicasting framework for wireless IP multimedia services, *Proc. ACM SIGMOBILE Workshop on Wireless Mobile Multimedia*, July 2001, pp. 13–23.
- D. Loguinov and H. Radha, Increase-decrease congestion control for real-time streaming: Scalability, *Proc. IEEE INFOCOM*, June 2002.
- S. Floyd et al., Equation-based congestion control for unicast applications, ACM SIGCOMM 30(4): 43–56 (Aug. 2000).
- D. Bansal and H. Balakrishnan, Binomial congestion control algorithms, *Proc. IEEE INFOCOM 2001* 2: 631–640 (April 2001).
- R. Talluri, Error-resilient video coding in the ISO MPEG-4 Standard, *IEEE Commun. Mag.* 2(6): 112-119 (June 1999).
- S.-J. Lee and S.-H. Jung, A survey of watermarking techniques applied to multimedia, *Proc. IEEE Int. Symp. Industrial Electronics 2001* 1: 272–277 (June 2001).

SURFACE ACOUSTIC WAVE FILTERS

PANKAJ K. DAS ROBERT J. FILKINS University of California at San Diego La Jolla, California

1. FUNDAMENTALS OF SAW DEVICES

Electronic systems have made use of acoustic waves for many years. Early examples of acoustoelectric devices include delay lines that exploit slow acoustic velocities to provide long delays in a small package, and high-Q filters that use quartz resonators. The use of surface acoustic wave (SAW) devices came about with the development of the interdigital transducer (IDT). The interdigital transducer allows SAW devices to be mass produced using IC fabrication techniques and thus reduces manufacturing costs tremendously. Filters in color television were the first consumer application of SAW devices. Today SAW filters play a very significant role in the wireless and cellular phone industries. A typical mobile phone uses half a dozen SAW filters, which might constitute one-fifth of the total fabrication costs. Because of the demands of the wireless industry, the center frequencies of SAW filters have been pushed from 2.5 to 5 GHz and beyond. The major objective of this article is to introduce the reader to the fundamentals of surface acoustic wave devices. A second objective is to discuss how these devices apply to communication systems, such as wireless transceivers, optical receivers, cellular phones, spreadspectrum processors, and RF filters in general.

Surface acoustic waves have been known to scientists since 1885, when Lord Rayleigh presented his paper to the Royal Society. For this reason, surface waves are often referred to as *Rayleigh waves*. The Rayleigh wave propagates along the surface of a solid material with particle motion in the plane defined by the surface normal and the propagation direction. The wave amplitude decreases with distance from the surface. The surface wave is quite strong when compared with a bulk wave as the energy spreads in only two dimensions instead of three.

SAW devices developed through the 1980s continued to utilize the Rayleigh wave in their operation. However, as the need for higher-frequency and lower-loss devices grew (with the onset of wireless communication needs), researchers began developing devices based on other types of surface waves. For example, so-called leaky surface waves (LSAW) have been used in the development of 900-MHz filters for wireless radio transceivers. Other quasisurface acoustic waves include surface skimming bulk waves (SSBW), surface transverse waves (STW), and Bluestein-Gulyaev (BG) waves. Devices employing these different surface or quasisurface waves look very similar. They all use IDTs for generation and detection. They differ in the crystal orientation of the substrate materials. Each type of device will be cut to a different orientation, leading to acoustic wave propagation along different crystal axes. The resulting pseudosurface waves have desirable properties, which we'll discuss later. We will generally refer to all these types of waves as surface acoustic waves.

Before describing surface acoustic waves in greater detail we shall make a digression into the use of bulk acoustic waves in the electronics industry. This provides a basis for understanding the later development of SAW devices.

1.1. Bulk Ultrasound Devices

Two basic types of ultrasonic waves can exist within a solid material, namely, longitudinal and transverse. In the longitudinal (or compressional) wave type, the displacement of particles within the material is in the



Figure 1. Bulk longitudinal ultrasound propagation in a isotropic solid [1].

direction of the propagating wave. As shown in Fig. 1, if a solid is divided into minute grid points, the longitudinal wave propagation causes some of the grid points to approach one another while other points are separated. The longitudinal wave moving in the z direction is composed of a series of compressions and rarefactions within the material. (Note that the lack of variation along the depth is the reason for calling a wave of this type a "bulk" wave.) Figure 2 shows the case of the transverse (or shear) wave type. The particle motion in this case is in the direction perpendicular to the direction of wave propagation, and the grid points now are translated either up or down. In three dimensions, there are two possible directions that are mutually perpendicular to the direction of motion. Therefore, in real solids two transverse wave modes are possible. These two shear waves are often referred to as vertically polarized (as in Fig. 2) and horizontally polarized (consider the particles in Fig. 2 were moving in and out of the page). The longitudinal and transverse bulk nondispersive waves can theoretically exist only within a solid of infinite dimension. For practical purposes, however, as long as the solid medium is considerably larger (e.g., 100 times greater) than the acoustic wavelength, the deviation from the ideal case is negligible.

The velocity of bulk transverse waves is roughly 3000 m/s, which is five orders of magnitude smaller than the velocity of electromagnetic waves in a solid — a fact that has been used to make ultrasonic delay lines. Consider the example of trying to obtain a $3-\mu$ s delay of a radar signal. To delay the electromagnetic wave, one requires 900 ft of coaxial cable. The equivalent delay for an ultrasonic wave traveling in solid substrate requires only 9 mm of material. The price to be paid for this compactness is the need for two transducers, one to convert the electrical energy to ultrasound, and a second to convert the ultrasound back to electrical energy. A scheme for a simple acoustic delay



Figure 2. Bulk transverse wave traveling in a solid.



Figure 3. (a) Simplified bulk longitudinal delay line of length L—delay time τ is equal to L/v, where v is the longitudinal acoustic velocity; (b) compact delay line using low loss, bulk shear waves [3].

line is shown in Fig. 3a. The transducers are generally made from a parallel slab of piezoelectric material such as quartz or lithium niobate. The parallel sides of the transducer are metallized so that voltages can be applied. The excitation voltage is an amplitude modulated signal of the form $h(t)e^{j\omega t}$, where $\omega = 2\pi f$ is the carrier frequency. The ultrasonic signal generated within the solid can be represented by

$$h\left(t - \frac{x}{v}\right)e^{j\omega(t - x/v)} = h\left(t - \frac{x}{v}\right)e^{j(\omega t - kx)}$$
(1)

where the wavenumber

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{v} \tag{2}$$

and v is the velocity of wave propagation. The output signal at the second transducer for a delay line of length l is given by

$$h\left(t - \frac{l}{v}\right)e^{j\omega t} \tag{3}$$

which excludes a constant factor that accounts for propagation or transduction losses. Here the delay time τ is given by

$$\tau = \frac{l}{v} \tag{4}$$

A small device of this type can provide long delays; such devices have been used in electronic systems since the 1940s. In fact, early devices also utilized shear waves and internal reflections to generate compact delay lines of the type shown in Fig. 3b. Two of the possible delay paths are shown in this figure. Many other examples can be found in early literature [3].

Of course, the device structure can be extended to provide multiple delays. The so-called tapped delay line is useful in many electronic applications such as radar



Figure 4. Simple representation of a tapped bulk wave ultrasound delay line.

processing. To obtain a tapped delay line with N taps using the abovementioned approach requires N + 1 transducers as shown in Fig. 4. In order to operate at high frequencies, each transducer must be carefully glued to the substrate, making the device rather cumbersome to fabricate.

A bulk wave in an unbounded medium is nondispersive. Often signal processing applications, such as pulse compression, require a dispersive delay, where each frequency travels with a different velocity. In this case a dispersive ultrasonic wave such as a plate or Lamb wave is used. A pulse compression filter, like those used in the 1950s to process radar chirp signals, is illustrated in Fig. 5. The dispersive property is shown in Fig. 6 as frequency versus delay time, which is



Figure 5. The use of a dispersive medium for pulse compression. The ultrasonic velocity is a function of the propagating frequency [1].



Figure 6. Dispersion represented by frequency as a function of time delay.

related to the frequency-velocity property of the wave. In other words, higher frequencies travel faster than do lower ones through the device. The desired input for this dispersive delay line is a "chirped" pulse whose instantaneous frequency is low at the beginning and increases progressively as function of time. As such a chirped pulse propagates through the delay line, the beginning of the wave is delayed more than the end of the wave. If everything is well matched, all the pulse energy will appear at the output at the same time, producing a spike [5]. Of course, the dispersion and the signal must be precisely matched in order to obtain good pulse compression.

1.2. Surface Acoustic Waves

Surface acoustic waves (SAWs) are more complicated than their bulk wave counterparts. The pure SAW or Rayleigh wave is a combination of the longitudinal and transverse wave components, which are elliptically polarized [1]; that is, the individual particles of the medium move in elliptical paths around their rest positions. This elliptical path is confined to the plane defined by the surface normal and the direction of wave propagation. A gridline representation of the particle behavior is shown in Fig. 7.

The exact derivation of the surface acoustic wave and how it comes to be confined to the free surface is quite complex [2]. The Rayleigh wave is a combination of the more general Lamb wave modes for a free plate. As the thickness of a plate becomes "infinite," the lowest-order antisymmetric (flexural) mode and the lowest-order symmetric (dilational) mode become degenerate, and combine to form the Rayleigh wave. The wave becomes tightly bound to the surface, leaving the interior of the plate undisturbed. The existence of such a wave was predicted from seismological research, where the earth's crust was considered to be a plate of infinite thickness.

The maximum SAW amplitude occurs right at the surface of the device. The wave amplitude decays rapidly in the direction perpendicular to the surface. The decay constant is approximately one wavelength long. For example, the acoustic wavelength in Y-cut LiNbO₃ at 100 MHz is roughly 36 μ m for a wave propagating in



Figure 7. The displacements of a rectangular grid of material points as a result of SAW propagation in an isotropic material.

2444 SURFACE ACOUSTIC WAVE FILTERS

the z direction. (Note that "Y-cut" means that the normal to the surface is the y-direction of the crystal.) Therefore the substrate of a 100-MHz SAW delay line need only be a few hundred micrometers thick in order to work. Figure 8a summarizes the properties of SAW, showing the compressional and shear component motions of the grid points as well as the stress depth in wavelengths. That the wave is confined to the surface means that energy will spread in only two dimensions, and attenuate as $1/r^2$, instead of $1/r^3$, as for the bulk wave case. This feature gives Rayleigh waves another advantage for delay-line devices.

The SAW velocity is approximately 90% of the shear wave velocity. This is an important point to consider. Since the SAW phase velocity is slower than lowest bulk wave velocity, the Rayleigh wavelength measured along the surface is larger than any projected bulk waves. The Rayleigh wave cannot, in general, phase-match to any bulk wave components. Only in the presence of certain types of anisotropy or surface discontinuities can SAW combine with bulk waves.

1.3. Pseudo-SAW or Shallow Bulk Waves

In addition to Rayleigh waves, SAW devices may exploit other modes of surface wave propagation. Leaky SAW (LSAW), surface skimming bulk wave (SSBW), and surface transverse wave (STW) modes can be produced using interdigital transducers on various engineered substrates. The main advantages of these devices are

- 1. Wave propagation velocities are much higher than pure Rayleigh waves, allowing devices to operate at higher frequencies for the same lithographic tolerances. That is, the spacing between IDT fingers will represent a quarter wavelength at a higher acoustic frequency.
- 2. The electromechanical coupling coefficient, K, can be larger, leading to an increase in operational bandwidth and lower obtainable insertion loss.
- 3. Pseudosurface waves penetrate deeper into the substrate, allowing for larger acoustic power handling before nonlinear piezoelectric and acoustoelastic effects occur.
- 4. The temperature coefficient of the acoustic velocity can be carefully tuned for pseudosurface waves to enhance stability.

A pictorial summary of the various surface acoustic wave modes is given in Fig. 8b.

1.4. Piezoelectricity

An ultrasonic wave moving through a solid is a manifestation of stress applied to a crystal lattice. The



Figure 8. (a) Representation of surface acoustic wave. The grid points go through both shear and compressional motions [6]; (b) summary of surface acoustic varieties: (1) Rayleigh or pure SAW mode; (2) leaky SAW propagation from surface into substrate; (3) surface skimming bulk wave (SSBW); (4) surface transverse wave (STW) [4].

resulting strain alters the ionic equilibrium positions for the lattice, and in some cases a net electric polarization occurs. This stress-induced polarization is known as the *piezoelectric effect*. Piezoelectric materials also exhibit the reciprocal behavior; when an electric field is applied to the crystal, the ionic positions move, the lattice changes size or shape, and strain is produced. This behavior allows piezoelectric materials to behave as acoustoelectric transducers.

The topic of piezoelectricity is treated in great depth in a number of other texts [3], but we shall briefly describe some of the key aspects of piezoelectric materials. First, what is it that makes a material piezoelectric? To answer that question, one should first consider that crystalline materials are comprised of a regularly repeating pattern of atoms. As a result of chemical bonding, the atoms in the lattice share electrons. When atoms of different sizes share bonding electrons, it is possible that the electrons will "spend slightly more time" with one of the elements. Therefore, on average, one atom can appear negatively charged, and another can appear positively charged. A net electric dipole is formed by the atoms. The polarization effect is accentuated by different interatomic distances, atomic sizes, and crystal orientation. This built-in dipole will repeat over the entire crystal, compensating itself so that the crystal has no net potential. At the ends of the crystal, impurities or other matter will eventually appear to compensate the built-in electric field. The application of mechanical stress will disturb the compensated dipole and create an electric potential. Alternatively, if an electric field is applied, the built-in dipoles will respond. The strength of the response depends on the orientation of the dipole to the applied electric field or mechanical stress. For example, the torque applied to a dipole is greatest when the electric field is initially oriented perpendicular to it.

What we have just described is a coupling between the dielectric and elastic properties of an anisotropic crystal. The precise coupling is described mathematically in terms of the elastic, dielectric, and piezoelectric constants. Their derivation is based on thermodynamic potentials and is beyond the scope of this article, but the key points are summarized. Each set of constants comprises a vector or tensor matrix. The elastic constants relate two secondorder tensors (stress and strain) and are therefore a fourth-order tensor. It can be shown that in general there are at most 21 (twenty-one) elastic constants for a crystal. The dielectric constants relate two vectors (electric and polarization fields) and is therefore a second-order tensor. There can be at most six (6) permittivities. The piezoelectric constants relate a second-order symmetric tensor (stress or strain) to a vector (electric field) and is therefore a third-order tensor. There can be as many as 18 (eighteen) piezoelectric constants. The electromechanical coupling efficiency, K (or sometimes K^2), is defined as the ratio of the mutual elastic and dielectric energy density (U_m) to the geometric mean of the dielectric (U_d) and elastic (U_e) self-energy densities.

$$K = \frac{U_m}{(U_d \times U_e)^{1/2}} \tag{5}$$

The value of K can also be described in terms of the individual material constants: s (elastic), ε (dielectric), and d (piezoelectric). A simple expression is

$$K = \frac{d}{(s\varepsilon)^{1/2}} \tag{6}$$

One final point to make about the effect of piezoelectricity on substrates is that it alters the mechanical stiffness of the material. The additional reaction force provided by the piezoelectric field can be combined with the normal elastic coefficients to create a set of stiffened elastic coefficients for the material [2]. The importance of this fact will be seen when we consider waveguides, transducers, and gratings.

In summary, we see that piezoelectric properties of a material are based on the absence of crystal symmetry, and that the exact piezoelectric behavior depends very much on the orientation of the crystal lattice to the applied external forces. These facts form the basis of substrate engineering for SAW devices.

1.5. Substrate Materials

A good piezoelectric substrate is a vital ingredient for a SAW device. Quartz is the only naturally occurring piezoelectric material used to manufacture devices. The largest piezoelectric constant for quartz is $d_{11} =$ $2.31(\times 10^{-12}C/N)$, which is smaller than other synthetic (human-made) SAW substrates [3]. Quartz does have the advantage of low acoustic loss, and low dielectric constant (which leads to low capacitance per unit length) when compared to synthetic piezoelectrics. Most of the crystals used in manufacturing devices are grown in a laboratory environment. Computer modeling of the crystal parameters was used to develop ST-cut quartz, the most readily used quartz cut for SAW devices. The object of STcut quartz is to provide very stable acoustic velocity over temperature; the temperature coefficient is nearly zero ppm/°C. The SAW propagation direction for ST quartz is the *x* direction with a velocity of $3.158 \text{ mm/}\mu\text{s}$.

Lithium niobate is probably the most important synthetic material for SAW substrates. It was discovered in the late 1950s, a time of great interest in synthetic piezoelectric materials. Typically, lithium niobate (LiNbO₃) used for devices is a congruent crystalline mixture of Li₂O and Nb₂O₅, composed of 48.6% Li₂O [22]. Most SAW devices are built of Y-cut, z-propagating LiNbO₃ substrates. The Rayleigh velocity is 3.487 mm/µs for this cut.

Lithium tantalate (LiTaO₃) is another popular substrate material. A rotated Y-cut, z-propagating crystal is used for pure SAW devices. The velocity is $3.254 \text{ mm/}\mu \text{s}$. The mechanical coupling factor is lower than that of lithium niobate, owing in part to a lower dielectric constant. However, lithium tantalate has the advantages of reduced degree of dielectric anisotropy, smaller temperature coefficient, and lower acoustic diffraction.

As described earlier, pseudo-SAW devices are obtained by using different crystal orientations and in some cases different substrate materials. Examples of useful orientations for propagation of LSAW on lithium niobate include 64° YX-cut, 41° YX-cut, and 36° YX-cut. Lithium tantalate is used as a substrate material for LSAW- and SSBW-type devices. For example, impedance filter elements based on SAW gratings or resonator structures employ a 36° YX-cut LiTaO₃ substrate. Each substrate and crystal cut has a different acoustic velocity, temperature coefficient, electromechanical coupling, and attenuation factor. The designer must choose what is best for a given filter application. The interested reader should consult Ref. 23 for a complete list of SAW substrates and their properties.

1.6. Thin Films

Thin films are often used in the design of surface acoustic wave devices [16]. As we shall see in subsequent sections, thin metal films are essential for the operation of SAW devices. They provide a means of generation, detection, and directional control of surface acoustic waves. The need for higher-performance devices pushes research toward better metal films. For example, to increase operating power and bandwidth requires metal films that resist electromigration and acoustic migration breakdown, while maintaining high conductivity at high frequencies.

Dielectric films can be added to SAW substrates to provide a variety of performance enhancements. A thin amorphous film, such as glass, deposited on a piezoelectric substrate provides the following advantages: surface passivation, reduction of pyroelectric effects (the buildup of a surface potential due to temperature changes), and smoothing to reduce propagation loss. Dielectric films are also used to modify the electromechanical coupling factor, reduce the frequency dependent temperature coefficient of velocity, and tune the acoustic velocity. Velocity tuning is an essential part of designing acoustic waveguides.

Piezoelectric films are used in several ways in advanced SAW devices. The first application is to provide substrate materials with high acoustic velocity. The higher velocity allows fabrication of higher-frequency devices without changing the lithographic feature sizes. Piezoelectric films are also useful for increasing coupling factors, and increasing the acoustic nonlinearity of the substrate. The nonlinearity is used in acoustoelectric convolver devices. A final advantage of piezoelectric thin films is to allow the integration of SAW devices with other microelectronic circuits. For example, depositing ZnO on silicon or gallium arsenide substrates to create integrated amplifier and SAW filter circuits.

2. SAW BUILDING BLOCKS AND DEVICES

2.1. Acoustic Impedance, Waveguides, and Gratings [2,5]

In order to understand how surface acoustic wave devices operate, we should say a little bit about wave reflection and transmission at material discontinuities. The Rayleigh phase velocity is the simplest parameter to describe the dispersion relation for the propagation of acoustic waves across the piezoelectric substrate. But in order to describe the behavior at material interfaces, a second parameter is often introduced: the *acoustic impedance*. The surface acoustic impedance is defined as

$$Z = \rho V_R \tag{7}$$

where ρ is the material density in kg/m³ and V_R is the Rayleigh wave phase velocity in m/s. This impedance definition allows one to adapt a transmission-line model for the wave propagation. At the interface between materials of different acoustic impedance, some of the acoustic energy is transmitted, some reflected. A reflection coefficient R and a transmission coefficient T can be defined as

$$R = \frac{Z_1 - Z_2}{Z_1 + Z_2} = \frac{\rho V_{R1} - \rho_2 V_{R2}}{\rho V_{R1} + \rho_2 V_{R2}}$$
(8a)

$$T = 1 - R \tag{8b}$$

This description is oversimplified, as it neglects the effects of wave polarization, incidence angle, and propagating modes, but it gives the reader a feel for what happens in a SAW substrate when the acoustic velocity changes abruptly.

As mentioned in the preceding section, the presence of thin films on the surface of a SAW substrate affects wave velocity. The first, and maybe most important, example is the presence of a metal film on the substrate. The conductivity of the metal film has the effect of shorting (short-circuiting) the piezoelectric field in the material. This, in turn, alters the stiffness and therefore the velocity of the material under the metal film. The change in SAW velocity due to the shorting of the piezoelectric field allows for direct measurement of the electromechanical coupling coefficient

$$K^2 = \frac{-2\Delta V_R}{V_R} \tag{9}$$

where ΔV_R is the fractional change in SAW velocity and V_R is the free surface SAW velocity. An impedance mismatch also occurs for the propagating surface wave. This action forms the basis of acoustic waveguides and gratings, which we will discuss next.

The basic function of an acoustic waveguide is to confine the acoustic wave propagation to an area of the substrate, much like a fiberoptic cable for light, or a coaxial transmission line for RF energy. The waveguide confinement is used in SAW devices to overcome beamspreading losses, to redirect signals, to correct phase fronts, to increase the acoustic energy density (important in nonlinear devices like convolvers), and to generally create the notion of an acoustic "circuit." All acoustic waveguides operate on the basis of the $\Delta V_R/V_R$ action, but different structures are possible to achieve this result. Waveguides can generally be divided into three main classes: flat overlay waveguides, topographic waveguides, and other engineered waveguides. Several possible structures exists within each class. Examples of each type are shown in Fig. 9.

The first type of waveguide we shall discuss is the flat overlay waveguide. The simplest type of overlay waveguide is called a strip waveguide (see Fig. 9a). In this waveguide, a thin layer of dielectric material is deposited on the SAW substrate. The dielectric strip will typically have a slower acoustic velocity when compared to the SAW substrate. The $\Delta V_R/V_R$ action confines the acoustic wave to within the slower material. Note that in the most general case,



Figure 9. Various types of acoustic waveguides structures. Flat overlay types: (a) mass loading strip of dielectric film; (b) shorting type with thin metal layer; (c) slot waveguide using shorting type metal films. Topographic waveguides; (d) rectangular ridge waveguide; (e) wedge waveguide, Engineered type; (f) in-diffused waveguide structure [5]; (g) acoustic grating structure.

neither the substrate nor the overlay strip needs to be piezoelectric.

A second type of overlay waveguide is the shorting (short-circuiting) strip guide. In this structure (Fig. 9b) a thin conducting film is placed on a piezoelectric SAW substrate. The conducting film short-circuits the piezoelectric field and reduces the acoustic velocity under the strip. The acoustic wave is weakly confined to the area under the strip. By changing the thickness of the conducting film, it is possible to introduce a mass loading effect as well. The combination of mass loading and piezoelectric shorting can be used to carefully tune the dispersion behavior of the waveguide.

A slot waveguide is produced by depositing a film with a faster acoustic velocity on the SAW substrate as shown in Fig. 9c. In this case the acoustic wave is confined to the "slot" area between the film overlays.

The topographic waveguide is produced by selectively removing an area of the substrate to create a ridge or wedge confinement region (see Fig. 9d,e). In the ridge or rectangular waveguide, the acoustic wave propagates as a function of the bending modes of the ridge section. The bandwidth of such a structure is therefore limited by the geometry of the ridge. The wedge is a means of creating a broader band waveguide. The tapered geometry of the wedge creates a wide range of frequencies that can propagate [5]. A key advantage of the topographic waveguide structure is its low loss.

One final type of waveguide structure is shown in Fig. 9f. This is the diffused waveguide structure. The acoustic velocity is altered within a region of the substrate by diffusing another material into the substrate. One example is to diffuse titanium into lithium niobate to create a confining region. The advantage of diffused waveguides over strip types is also lower loss.

Another SAW device structure based on the $\Delta V_R/V_R$ effect is the acoustic grating. The object of the grating is to create an acoustic reflector. The reflecting element can then be used to create a variety of wave control

devices including filters. The device is analogous to the Bragg grating in optics. The grating structure, as shown in Fig. 9g, is a series of strips. The strips may be created using any of the structures described for waveguides. Two of the more common structures are shorting metallic strips and ridge guides. The metallic strips are deposited and etched using photolithography, whereas a ridge structure can be created by e-beam (electron-beam) lithography. The strips are separated by a half the acoustic wavelength of interest. Recall that at each impedance interface, an amount of the energy is reflected. Separating the strips by one-half wavelength $(\lambda/2)$ allows each reflection to add in phase. A large number of strips in sequence are needed to reflect all the energy. The optimum number of strips is a function of the structure used and the substrate material. For example, to obtain near-100% reflection using shorting strips on lithium niobate requires on the order of 100 strips [4].

This brief introduction into acoustic waveguides and gratings lays the foundation for discussing the structure and design SAW devices, the devices that, in turn, are used to create SAW filters. The $\Delta V_R/V_R$ effects in some cases are the basis of device operation, and in other cases (such as IDTs) create nonideal behaviors that must be compensated for in some way. Now let's discuss interdigital transducers.

2.2. Interdigital Transducers

The interdigital transducer (IDT) is shown in Fig. 10b. The IDT is simply a set of metal strips placed on the piezoelectric substrate. Alternate strips, or "fingers", are interconnected to form two interdigitated electrical contacts. The width of each finger and the distance between fingers is usually one-quarter of the acoustic wavelength to be generated. When a RF voltage is applied to the two contacts, an electric field is simultaneously set up between all adjacent fingers. This has the effect of alternately compressing and elongating the piezoelectric substrate between the fingers, producing a distributed



Figure 10. Generation of a surface wave using a bulk transducer—this is improved using IDTs: (**a**) SAW delay line made from two IDTs; (**b**) diagram of an IDT composed of a set of metal strips placed on a piezoelectric substrate [7].

source of strain. The resulting acoustic waves propagate to the right and left, as the superposition of all of the waveguide modes of the substrate plate structure. Because the electric and strain fields are confined close to the substrate surface, the generation of surface waves is strongly favored. The type of surface waves excited are a function of the applied electric field, the piezoelectric matrix for the substrate, and the relative orientation of the crystal axes. In some cases, for example Y-cut lithium niobate, Rayleigh waves are generated. In other cases, for example, 36° YX-cut lithium tantalate, LSAW waves are favored.

To efficiently generate a surface acoustic wave, all the sources must add coherently. Consider the RF voltage with period T; to propagate a distance $\lambda/2$ takes T/2 seconds. This is also the time it takes for the electric field between two adjacent fingers to reverse polarity. Therefore, all the generated waves will add together constructively. The reverse process occurs for detection; that is, the elastic deformation passes under the fingers and induces in-phase voltages between each pair of electrodes. This type of operation makes the IDT a simple and efficient SAW generator.

Next, we shall describe the operation of SAW delay lines, and in the process we'll uncover some of the limitations of and improvements for the simple IDT structure.

3. SAW DEVICES FOR FILTERS

3.1. Delay Lines

A delay line is constructed using two IDTs on a piezoelectric substrate as shown in Fig. 10a. If lithium niobate is used as the substrate the surface velocity is about 3600 m/s, which translates into an acoustic wavelength of 36 μ m at 100-MHz operation. Using the guidelines previously mentioned, the IDT finger width and spacing would be 9 μ m. Of course, the delay is a function of the spacing between the two IDTs; For example, if the IDTs are separated by 1 cm, the delay for the SAW device will be 2.3 μ s.

Why use many finger pairs instead of only one pair for IDTs? Actually, an IDT can be (and sometimes is) made using only a single pair. However, greater transducer efficiency is achieved for the same applied voltage, since for N finger pairs the resulting N sources will add coherently. As if often the case, there is an inherent gain-bandwidth trade-off. There is a reduction in operating bandwidth as more and more finger pairs are added. In general, the optimum number of finger pairs (N_{opt}) is proportional to the reciprocal of the electromechanical coupling coefficient [5]. That is

$$N_{\rm opt}^2 = \frac{\pi}{4} \frac{1}{K^2}$$
(10)

where K^2 is the squared coupling coefficient of Eq. (9).

The two characteristics of SAW delay line filters that limit the performance are delay-line losses and tripletransit echo. Delay line losses can be divided into four: IDT loss, propagation loss, misalignment or steering loss, and diffraction loss.

An inherent loss of 6 dB occurs in a SAW delay line because an ordinary IDT generates waves in the forward and backward directions. This division of energy gives 3 dB of loss per IDT, transmitter, and receiver. The loss can be reduced or eliminated through the use of unidirectional transducers. The design of unidirectional transducers is reviewed in the next section.

Now let's move to the second loss mechanism in SAW delay lines — propagation loss. Propagation loss is characterized by a frequency-dependent loss per unit length. The following functional form can be used to describe this behavior:

$$A(x) = A_0 e^{-a(f)x}$$
(11)

As SAW propagates along the surface of the device, the amplitude starts out with amplitude A_0 , and decays as function of x and $\alpha(f)$. As the SAW frequency increases, the material deformation is unable to accurately follow the RF field and produces a phase term. This, in turn, creates a propagation loss. In general, the value of $\alpha(f)$ is proportional to the square of the frequency. An additional loss term is created by air loading of the surface. If the delay line operated in a vacuum, the loading loss would be zero. In most cases, a SAW device package is hermetically sealed with an inert gas such as nitrogen. In this case the loading loss is negligible, becoming more dominant at low frequencies.

Beam steering loss is ideally zero if the substrate is cut to the proper direction. The proper direction is the one where the power flow angle is zero. If the IDTs become misaligned with the crystal axis, or the cut is incorrect, the power flow angle will be nonzero, and loss will develop. As with propagation loss, the steering loss is worse for long delay lines.

Diffraction loss is the result of the finite-sized aperture of the IDT. The acoustic wavefront can be described in terms of a near-field, or Fresnel region, and far-field or Fraunhofer region. This is just as in the case of optics. In the far-field region the wavefront spreads out to a width larger than the aperture of either the transmitting or receiving IDT, placing a limit on the length of a delay line for a given IDT aperture length. The calculation of this loss is complicated by the anisotropic nature of the substrate crystals. Models are available that predict this loss to within a fraction of a decibel. In principle, acoustic waveguides can be used to overcome this loss in long-delay devices.

3.1.1. Triple-Transit Echo. We expect that when an RF signal is applied to a SAW delay line, the output will be an exact copy of the input, delayed only by some time τ and perhaps lower in amplitude. It is possible for other outputs to occur. The most important and troublesome extraneous output is the triple-transit echo. As the name suggests, the extra output appears at time $t = 3\tau$ and is caused by reflections of the SAW off the IDTs. For a delay line with two matched IDTs, this echo is 12 dB lower than the main signal. It is worth noting how so much SAW energy can be reflected off the IDTs. As mentioned previously, the acoustic velocity is different in the regions of the substrate covered by metal electrodes. The net $\Delta v/v$ change in the acoustic impedance along the SAW transmission line generates small reflections at each IDT finger. The reflections add coherently due to phase matching, and the result is a significant extra echo. A split-finger IDT is one means of reducing the phase-match condition from occurring (see Fig. 11). Each quarter-wave finger pair is separated into two lambda (2λ) by eight fingers. A multistrip coupler can also be used to trap the triple transit echo (this will be discussed in a subsequent section).

If we pause here for a moment and compare a SAW delay line to a bulk ultrasound equivalent device, we immediately see the advantage of the SAW device. The SAW device can be fabricated using a one-step photolithographic process, whereas the bulk ultrasound device required gluing at least two transducers to the substrate. In fact, for a delay line with N taps, the bulk ultrasound device would require gluing N+1 transducers. The SAW device is fabricated in one step regardless of the number of taps, owing to the lithographic processing. The cost advantages in both time saved and device yield are clear. Also, the two-dimensional nature of the SAW device makes it consistent with other integrated circuit techniques for packaging and testing.

3.2. Unidirectional Transducers

Unidirectional transducers correct the 3-dB loss of the bidirectional IDT, and overcome the problem of triple-transit echo. There are four techniques for making low-loss unidirectional transducers:

- 1. Two IDTs separated by a quarter wavelength
- 2. Three-phase excitation
- 3. Single-phase unidirectional transducer (SPUDT)
- 4. Multistrip couplers

The first type of *unidirectional transducer* is shown in Fig. 12. The transducer consists of two IDTs separated



Figure 11. Diagram of transducers with (a) single electrode and (b) double or split electrodes per half wavelength. The split electrodes provide a means of canceling the triple-transit echo in delay lines.



Figure 12. A unidirectional transducer using a reflector IDT spaced a quarter wavelength from the driven IDT [9].

by a quarter-wavelength gap, which produces a phase difference of 90° . The second IDT is also terminated such that the load will completely reflect the SAW. In operation, the excited backward wave is reflected by the second



Figure 13. (a) Three-phase excitation of an IDT [10]; (b) the resulting interelectrode electric field evolution for three-phase excitation.

transducer with a phase shift, allowing it to reinforce the forward moving wave. This cancels the 3-dB loss of a simple IDT. One drawback, however, is a limitation on achievable frequency response.

The three-phase excitation IDT is more complex, but effective. As shown in Fig. 13, the IDT now consists of three electrode fingers. The RF drive signal is converted into a three-phase signal using a 60° phase shifter. Each drive signal is thus separated by 120° , as illustrated in Fig. 13a. The resulting interelectrode electric field evolution is shown in Fig. 13b. Only the forward traveling wave propagates; the backward wave is canceled out. Of course, the direction can be reversed by adding an additional 120° phase shift to the drive signals. Fabrication of the three-phase IDT is further complicated by the need for an extra insulator to isolate one of the IDT fingers.

The single-phase unidirectional transducer (SPUDT) is shown in several forms in Fig. 14. The SPUDT uses floating electrodes arranged to form an acoustic grating, designed to reflect the traveling SAW. When properly designed, the reflector reinforces the forward moving SAW, and cancels the backward moving SAW. The basic SPUDT device has three configurations [4]: single floating electrode (Fig. 14a), double floating electrode (Fig. 14b), and a comb type (Fig. 14c).



Figure 14. Finger layout for a single-phase unidirectional transducer (SPUDT). The basic device has three configurations: (**a**) single floating electrode; (**b**) double floating electrode; (**c**) comb-type SPUDT using longer grating elements [2].

3.3. Surface Acoustic Wave Filters Based on Delay Lines and Apodization

We shall now begin our discussion of SAW filters by showing how a tapped-delay structure can be used to create a transversal, or finite-impulse-response (FIR), filter. The concept of transversal filters has been developed fully in other texts. We start here by noting that the transversal filter has a transfer function given by

$$H(\omega) = \sum_{n=0}^{N-1} C_n e^{-j\omega T_d}$$
(12)

This transfer function corresponds to a delay line where each of the N output taps is multiplied electronically by the appropriate coefficient C_n . Each of the N products is then summed electronically. A programmable filter can be achieved by simply switching different values for $\{C_n\}$. If one is merely interested in a fixed transversal filter, the value of C_n can be designed into the transducer itself. The transducer gain adjustment can be achieved many different ways, but we shall discuss a method called *apodization*.

Consider the IDT shown in Fig. 15 and assume that the same SAW is incident on both finger pairs A and B. One sees that the finger pair A overlaps for the entire width of the SAW wavefront as drawn. The finger pair B, however, overlaps for only a 10% fraction of the wavefront region. It is therefore expected that the voltage detected for each transducer will correspond to coefficients $C_A = 1.0$ and $C_B = 0.1$. SAW devices become an easy platform for making inexpensive filters. The calculated coefficient values are incorporated directly into the mask used to fabricate the IDT pattern. For example, to implement some specific transfer function $H_1(\omega)$ [assuming that $H_1(\omega)$ is suitably band-limited], one can expand $H_1(\omega)$ in a Fourier series to generate a series as in Eq. (12). In particular, if the impulse response of the filter is $h_1(t)$ then the coefficients $\{C_n\}$ are given by

$$C_n = h_1(nT_d) \tag{13}$$

where T_d is now chosen to provide the correct sampling interval for reconstruction of $h_1(t)$. When we look at the SAW tapped delay line under a microscope, we actually see a spatially sampled replica of the impulse response in the IDT finger pair overlap. Any of the wellknown techniques for digital filter design can be applied



Figure 15. Overlap of the IDT finger pairs determines the transversal filter coefficients [7].

to designing SAW filter masks using apodization. The production of finite impulse response or transversal filters is readily achievable.

Now let's examine the frequency response of the basic IDT arrangement using the idea of a tapped delay line with N fingers. In the simple case all the IDT finger pairs are the same, which corresponds to all $C_n = 1$ using the notation given above. The spatially sampled impulse response is therefore a rectangular pulse. The frequency-domain transfer function for this arrangement is

$$H(\omega) = \sum_{n=0}^{N-1} e^{-j\omega T_d} \qquad \text{where} \quad T_d = \frac{1}{f_0} \tag{14}$$

and f_0 is the center frequency of the filter design. In the vicinity of f_0 , the magnitude of the transfer function (14) is approximately

$$|H(\omega)| = N \frac{\sin(N\pi \Delta f/f_0)}{N\pi \Delta f/f_0} \quad \text{where} \quad \Delta f = f - f0$$

and $f = \frac{\omega}{2\pi}$ (15)

This is the familiar sinc function response centered around f_0 , which has bandwidth that is inversely proportional to N. N in this case is the spatial representation of time; a larger N means a longer impulse or time response for the filter. Thus the bandwidth of the filter shrinks as the number of IDT pairs is increased. That the frequency response is the Fourier transform of the aperture or apodization is a very satisfying and useful result for understanding SAW filters. Figure 16 summarizes the Fourier transform behavior of the SAW filter.

When selecting SAW filter bandwidth, one must also consider the device insertion loss. Figure 17 shows the fractional bandwidth versus minimum theoretical insertion loss of a SAW delay line using different materials. The minimum insertion loss is 6 dB and increases as the fractional bandwidth of the filter is increased. (The plot assumes that the IDTs are bidirectional.) Using unidirectional transducers, it is possible to build a SAW filter with 0 dB insertion loss. The optimum fractional bandwidth is represented by the expression

$$\left(\frac{\Delta f}{f}\right)_{\rm opt} = \frac{1}{N_{\rm opt}} = \frac{2K}{\pi^{1/2}} \tag{16}$$

where N_{opt} is the optimum number of finger pairs and K is electromechanical coupling coefficient.

The apodized IDT (see Figs. 18–20) in its simplest form has the problem of increased diffraction loss. This is due to variations in electrode overlap. When the electrodes overlap, only a small amount a large nonmetallized area results. As the SAW passes under the apodized electrodes, it sees a varying, parasitic acoustic grating. The result is wavefront distortion and added diffraction losses. An improved IDT, shown in Fig. 18, corrects this problem by adding floating electrodes. The floating electrodes maintain a consistent grating structure. tics [7].

70

60

Insertion loss (dB)

20

10

6 0

1



Figure 17. Minimum theoretical insertion loss versus fractional bandwidth for different piezoelectric substrates used for SAW devices.



Figure 18. Improved IDT for transversal filter applications. This structure uses split-finger IDT to reduce triple-transit echo, and floating electrodes (gray stripes) to reduce the diffraction effects of apodized electrodes.

Figure 19. Typical response of a SAW filter using apodization [11].

3.4. Multistrip Couplers

The multistrip coupler (MSC) is a useful structure for creating unidirectional IDTs and for suppressing undesirable bulk modes in SAW filters. In this section, we shall develop a simple understanding of the multistrip coupler and discuss some of its applications. The multistrip coupler is a set of parallel metal fingers that are not electrically connected to each other. To appreciate the operation of these fingers, we must first consider the two-dimensional wavefront of the propagating surface acoustic wave.

3.4.1. Simple Theory. A typical multistrip coupler is shown in Fig. 21. The device is comprised of two pairs of



Figure 20. SAW frequency response parameters [7].



Figure 21. Multistrip directional coupler with IDTs at the input and output ports [8].

IDTs, labeled A_1, A_2 , and B_1, B_2 . The normal path for SAW would be from A_1 to A_2 , and from B_1 to B_2 , thus the device has two acoustic paths or tracks. We shall call these tracks A and B. The multistrip coupler is formed by the parallel rows of fingers that cross the device, extending over both track A and track B. With the addition of the multistrip coupler, the acoustic path is altered such that if the IDT A_1 is excited, then all the acoustic energy is transferred to track B and appears at IDT B_2 . The output at A_2 will be zero if the coupler is properly designed. If only IDT B_1 is excited, then the SAW energy is transferred to track A and appears at IDT A_2 . At first glance, it might seem impossible for the mechanical energy of the SAW to be transferred from one track to another simply by virtue of the parallel metal fingers. However, it is important to recognize that the piezoelectric substrate produces an electric field and thereby voltages on the metal fingers as the acoustic wave passes under them. This provides a means of coupling a common electric field between the two tracks. The electric field, in turn, produces a surface acoustic wave in the previously unexcited acoustic track. By using a multitude of parallel fingers, this passively generated SAW can be made quite strong.

The presence of metal fingers in both of the tracks on the piezoelectric substrate leads to coupling between the two tracks. This provides a means for energy transfer for a long multistrip coupler (MSC). The operation of the MSC and the coupling effect can be understood by considering Fig. 22. In this figure the SAW wavefront generated by the transducer in track A has been separated into symmetric and antisymmetric parts, labeled s and a, respectively. The wavefront of the symmetric part extends uniformly over both tracks and has a magnitude of A/2, where A is the amplitude of the total SAW wavefront. The antisymmetric part has the same A/2 magnitude and extends over both tracks; however, in track B the amplitude is 180° out of phase with respect to the amplitude in track A. The amplitude in track A has the same phase as the symmetric component. Without the metal fingers, if the symmetric and antisymmetric parts are combined, then the SAW cancels in track B and the full amplitude appears in track A.

While both the symmetric and antisymmetric parts propagate under the MSC, we note that there is an important difference between the two. The symmetric part of the SAW induces voltages on the fingers that are the same in both tracks. In contrast, the antisymmetric part of the SAW induces voltages that are of opposite phase on either ends of the metal fingers. Because the couplers are unable to support current flow parallel to the fingers, the antisymmetric field is essentially unaffected by the presence of the metal contact. In other words, the antisymmetric field does not "see" the metal contacts and propagates with the unstiffened velocity. The symmetric part of the wave will travel with the stiffened velocity. Well, actually because of the gaps between the metal fingers the symmetric and antisymmetric wavefronts will not be exactly equal to the stiffened and unstiffened velocities. The purely stiffened and unstiffened velocities correspond to the completely metallized and bare substrates respectively.

Using these arguments it can be shown that complete power transfer from track A to track B can take place for an MSC of length L_T

$$L_T = \frac{\lambda}{K^2} \tag{17}$$

where the coupling coefficient $K^2 \sim 2\Delta v/v$. If *d* is the MSC repeat distance, then the number of fingers N_T needed for



Figure 22. The input and output field distributions of a MSC separated into a symmetric (s) and antisymmetric (a) modes; the diagram shows 100% coupling from track A to track B [8].

a complete transfer is

$$N_T = \frac{\lambda}{K^2 d} \tag{18}$$

For a length *x* of the MSC, the power in each of the tracks is given by

$$P1 \sim \cos^2\left(\frac{\pi x}{L_T}\right)$$
 (19a)

$$P2 \sim \sin^2\left(\frac{\pi x}{L_T}\right)$$
 (19b)

The following is a brief list of possible applications for multistrip couplers:

- Coupler
- Bulk wave suppressor
- Coupling between different substrates
- Aperture transformations
- Precise attenuator phase correction by offset
- Delay-line tap
- Beamwidth compressor
- Magic tee
- Beam redirection
- Reflector
- Unidirectional transducer
- Reflecting track changer
- Echo trap
- Better filter design
- Multiplexing
- Strip coupled amplifier and convolver
- Compressed convolver

Details of the applications can be found in the literature [7,8].

3.5. SAW Oscillators and Resonators

There are two distinct ways in which a SAW device can be used as the high-Q resonator circuit in an oscillator. The first case is shown in Fig. 23, where a SAW delay line is used in the feedback path around an amplifier. The second case is a SAW resonator structure, a planar cavity formed by two grating reflectors, which results in the high-Q element. The SAW-based oscillator has many advantages over bulk wave quartz oscillators in the frequency range beyond 100 MHz up to \sim 5 GHz. The operating frequency of bulk wave devices is based on crystal thickness, and for these high-frequency applications the required thickness are too thin to fabricate reliably. This means that for higher-frequency devices, overtones of the bulk resonator's fundamental frequency must be generated, which, in turn, requires multipliers and associated filters. The SAW device does not require these multiplier/filter combinations, thereby providing great advantage in size, weight, power, and cost.

The delay-line oscillator is a phase shift oscillator that operates at a frequency of

$$f_n \approx n \frac{v}{L} \tag{20}$$

where *n* is an integer. Of course, in order to oscillate, the amplifier gain must compensate for the delay-line losses, and the phase shift around the feedback path must be multiples of 2π . A particular frequency can be selected by designing some filter properties into the IDTs used to constitute the delay line.

In SAW resonators, the acoustic wave is trapped in a planar cavity formed by two acoustic grating reflectors. (In a conventional bulk wave resonator, the acoustic wave is confined by the two parallel surfaces of the crystal. The acoustic impedance mismatch between the air and crystal form a perfect reflector.) As mentioned in Section 2.1 with respect to acoustic gratings, the SAW decomposes into reflected longitudinal and shear waves when incident on

VCC



Figure 23. (a) An oscillator using a SAW delay line to produce the required phase shift; (b) discrete Pierce oscillator using a SAW-based impedance element Y1.



 $\label{eq:states} \begin{array}{l} \mbox{Figure 24. (a) Typical one-port SAW resonator; (b) two-port SAW resonator.} \end{array}$

an abrupt surface discontinuity. The acoustic grating is made up of a large number of small, periodic surface perturbations.

The typical schematic for one-port and two-port SAW resonators is shown in Fig. 24. In a two-port resonator separate IDTs are used for generation and detection of the surface waves. The number of reflectors in the grating strips is typically in the hundreds. The performance of a resonator can be characterized almost entirely in terms of its reflectors. The Q factor of the resonant cavity with no propagation loss can be written as

$$Q = \frac{2\pi l}{\lambda (l - |r_f|^2)} \tag{21}$$

where l is the effective cavity length and $|r_f|$ is the amplitude reflection factor. The energy lost in the reflector due to mode conversion or absorption reduces the reflection factor and therefore the *Q*. The cavity length, *l*, is the sum of the separation between the gratings and the penetration of the SAW into the grating regions. The penetration of the SAW into the grating regions again depends on the design of the gratings. Oscillators with a Q of 30,000 have been reported using SAW resonators, compared to a Q of only a few thousand for oscillators based on delay lines. Using resonators, an oscillator can be made with stability and noise performance as good as in devices made from quartz overtone crystals, and with higher operating frequencies. For example, Rayleigh wave resonator oscillators on STcut quartz have a typical noise floor of -176 dBc/Hz with long-term stability on the order of 1 ppm per year [2]. For this reason, SAW resonators find use in UHF and VHF oscillators and wireless filters.

SAW resonators are a building block for filters operating in the 0.9–5-GHz range.

3.6. Convolvers

The nonlinear interaction of surface acoustic waves in materials can be used to make convolvers for signal processing. The nonlinearity of the material can be produced by either large-amplitude acoustic signals or by acoustoelectric interactions. Acousto-electric SAW convolvers can be configured in three different ways: separate medium, combined medium, and strip-coupled. What this classifications mean and how each operates will be discussed in the following section.

Let's begin with convolvers based on acoustic nonlinearity. The large amplitude acoustic nonlinearity can be considered the breakdown of Hooke's law within the material. One generally uses a piezoelectric substrate as the nonlinear acoustic material. This has two advantages: (1) piezoelectric nonlinearity is less than elastic stress-strain nonlinearity and (2) more importantly, the second harmonic of the strain wave has an associated electric field that is easy to detect and integrate by measuring the total voltage across (or current flowing through) the interaction region.

Convolvers based on acoustic nonlinearity have been demonstrated at different frequencies. They tend to have large inherent loss because the nonlinear coupling coefficient, K, is small. The dynamic range of these convolvers is also limited by failure of the elastic media. Higher power levels must be achieved in order for these devices to work well. Multistrip couplers, acoustic waveguides, or curved transducers are often used to boost power levels. Examples of each type of device structure are shown in Figs. 25-27.

Acoustoelectric phenomena produce a different type of nonlinearity. This nonlinearity is produced by the interaction of a SAW-generated piezoelectric field with the space charge arising from the displacement of free carriers in a semiconductor. The current density within the semiconductor is proportional to the product of the free-carrier density and the piezoelectric field. The carrier



Figure 25. Diagram of a convolver using multistrip coupler compression, acoustic waveguides.



Figure 26. Diagram of a convolver using horn waveguide to increase power density.



Figure 27. A convolver using curved transducers and a waveguide.

density is itself a function of the applied electric field, since the carriers are rearranged to produce a space charge. Thus, the normal component of the electric field at the semiconductor surface creates a voltage that is proportional to the square of the field. Two possibilities exist for creating an acoustoelectric convolver. One example is to use a piezoelectric semiconductor, such as GaAs, to generate the SAW — this is known as a *combined medium structure*. A second possibility is referred to as the *separate medium structure*, where a semiconductor is placed in close proximity to a piezoelectric substrate such as lithium niobate (Fig. 28). One can also evaporate the semiconductor directly over the piezoelectric substrate. A final variation is to evaporate a piezoelectric film such as ZnO over the semiconductor substrate.

3.7. Summary of SAW Devices

As late as the early 1980s SAW devices were a scientific curiosity; now they have matured to the point where they are routinely used in communication and signal processing applications. SAW filters can be divided into five categories:

- 1. Tapped delay line with fixed taps
- 2. Tapped delay line with programmable taps



Figure 28. Diagram of a convolver made using a separated medium structure [12].

- 3. Resonators
- 4. Convolvers
- 5. Transform-domain processors

Let's take a moment to summarize the key feature of each device category. A tapped delay line with fixedweight taps produces a filter with a very steep filter transition band; that is, when compared to an LC filter, the SAW delay-line filter produces larger stopband rejection closer to the corner frequency. A tapped delay line with programmable taps is used as programmable, matched, or with some added complexity as Widrow-Hoff LMS processor. Resonators can be used in oscillator circuits or cascaded to build narrow passband filters. A convolver is a three-terminal device, which utilizes its nonlinear response to perform programmable convolution of signals. Finally, transform-domain processors are subsystems used in real-time signal processing. For example, a realtime Fourier transform processor can be created using a SAW chirp filter.

4. SAW FILTER DESIGN

4.1. Finite-Impulse-Response Filters

In the previous section, we introduced the notion of a finiteimpulse-response (FIR) filter based on apodized IDTs. In reviewing the loss mechanism in delay line devices, improvements were made to the basic IDT to arrive at the structure shown at Fig. 18. A high-performance IDT uses split electrodes and floating electrodes to compensate for nonideal behavior. There are, in addition to insertion loss and fractional bandwidth, other parameters that one is concerned with in FIR filter design. These other parameters are listed in the table in Fig. 19, along with typical values achieved using SAW filters. The overall design parameters from the table in Fig. 19 for a filter are shown in Fig. 20 [7]. These include insertion loss, pass bandwidth, rejection bandwidth, transition bandwidth, fractional, bandwidth, rejection, shape factor, amplitude ripple, phase ripple, and group delay. Here again, the overall design of a filter becomes quite complicated and beyond the scope of this short article. We do wish, however, to highlight the performance that one can achieve by using SAW for FIR filters. One such high-performance filter is shown in the top trace in Fig. 19. The overall rejection of roughly 90 dB demonstrates the advantages of SAW devices over conventional *LC* technology.

4.2. SAW Resonator Filters

For many designs operating beyond 100 MHz, filters based on SAW resonators have been more advantageous. Indeed, for operating frequencies beyond 1 GHz, resonator structures have proved to be the best path for SAW designers. These designs provide low insertion loss, flat passband, steep transition, and excellent close-in rejection. Filters operating at 2 and 5 GHz based on LSAW impedance elements have been demonstrated [15]. The resonator structures for these filters have been etched into lithium tantalate substrates using electron-beam



Figure 29. General filter form for a Cauer-type ladder [14].

lithography. Feature sizes on the order of 200 nm are necessary to create 5-GHz acoustic gratings. The SAW impedance elements are designed into ladder networks, as would be used for conventional LC filter design. Let's review ladder networks.

The general Cauer form of a ladder network is shown in Fig. 29 [14]. The ladder is composed of series and shunt impedances Z_1 through Z_n . The process of synthesizing a filter response involves partial fraction expansion of the desired transfer function. Each term in the expanding transfer function corresponds to the impedance at each ladder rung. In practice, predetermined filter prototypes are used to create ladder filters. These prototypes include the maximally flat Butterworth, the equiripple Chebyshev, and the linear phase Bessel filter responses. The designer can look up the required filter coefficients (*L* and *C* values) for the type and order of the filter. Once a lowpass prototype design is established, the filter can be translated to the bandpass of interest by replacing the Ls and Cs with LC combinations. As shown if Fig. 30, standard equations are available to scale a design to the desired operating point.

The equivalent circuit model for a periodic SAW grating shown in Fig. 33 shows how it, too, forms a ladder network. All the SAW ladder elements, Z_1 , Z_2 , Z_{1m} , Z_{2m} , are functions of the electrode width, spacing, and acoustic velocity. With the aid of CAD, one can use the relations shown in Fig. 33 to synthesize impedances. Each SAW resonator can, in turn, be generalized to create a ladder network as shown in Fig. 31. The SAW impedance element filters (IEFs) are combined into ladder networks. The results are summarized in the literature [15,17].

A high-performance 900-MHz filter design, based on LSAW resonators, is shown in Fig. 32. The design is based on a two-port SAW resonator structure on 36° LiTaO₃. Two outer IDTs drive an interior and two exterior grating structures. The entire resonator is then duplicated in an image connection design technique [17] (Note how the filter image has horizontal and vertical symmetry). The



Figure 30. Designing ladder networks is often performed by first synthesizing a lowpass filter using LC pair (**a**), and then scaling the response using a bandpass transformation (**b**) [14].



Figure 31. Ladder network composed of SAW impedance element filters (IEF), for synthesizing filters in the gigahertz range [15].



Figure 32. Low-loss, sharp-cutoff 900-MHz ISM band SAW filter design [17]. Filter structure utilizes image connection design technique, combined with resonator structures to produce excellent performance. Critical design spacings, A and B, highlighted in diagram.

design procedure involves: selecting the grating period, then adjusting the IDT to grating spacings (labeled A and B in Fig. 32). Each spacing is adjusted to minimize ripple, and maximize the transition band. The overall frequency response is the combination of the IDTs by themselves, and the grating response. So the response of each is tuned to create the best overall response. The final filter specifications are 903 MHz center frequency, 5.6 MHz transition bandwidth, 2 MHz passband, 2.4 dB insertion loss, 65 dB sidelobe suppression, and a die size of 3 mm².

Other designs are shown in Figs. 33–35.

4.3. SAW CAD Design

Several computer-aided design tools exist for the development of SAW filters. One such example is SAWCAD-PC available online from the University of Central Florida (*http://www.ucf.edu*).

5. SAW FILTER APPLICATIONS IN TELECOMMUNICATIONS

The following is brief list of telecommunication applications where SAW filters are often employed:

- 1. Nyquist filters for digital radio
- 2. IF filters for mobile and wireless transceivers



Figure 33. Improved equivalent-circuit model for IDT design that considers secondary effects of stored energy and bulk waves [17].



Figure 34. General block diagram for a antenna duplexer in a wireless system. Two SAW- or LSAW-based filters are used to allow transmit and receive frequencies to share a single antenna.



Figure 35. A simple ladder-type filter structure used for designed Tx and Rx filters. Each element Z_s and Z_p is made from a SAW resonator structure [16].

- 3. Antenna duplexers
- 4. Clock recovery circuit for optical fiber communication
- 5. Delay lines for path length equalizers
- 6. RF front-end filters and channelization in mobile communications

- 7. Precision fixed frequency and tunable oscillators
- 8. Resonant filters for automotive keyless entry, garage door transmitter circuit, and medical alert transmitter circuit.
- 9. Pseudonoise (PN)-coded tapped delay line
- 10. Convolvers and correlators for spread-spectrum communications
- 11. Programmable and fixed matched filters.

A partial list of commercially available SAW filters is shown in Table 1. The table includes applications, center frequency, bandwidth, and typical insertion loss for each device. In the following sections, some of these applications will described in greater detail.

5.1. SAW Antenna Duplexers

An antenna duplexer is an example of the use of SAW filters in modern wireless applications. A general diagram of an antenna duplexer system is shown in Fig. 29. The transmit channel (Tx) outputs signals from a power amplifier (PA) and conveys them to the antenna via a transmit filter. The receive channel (Rx) decodes incoming signals by first processing them through a receive filter (Rx) and then amplifying with

	Center			
	Frequency		Fractional	Insertion
Application	(MHz)	$BW\left(MHz\right)$	BW (%)	$Loss\left(dB\right)$
900-MHz cordless phone Tx	909.3	3	0.33	5
900-MHz cordless phone Rx	920	3	0.33	5
AMPS/CDMA Rx	881.5	25	2.84	3.5
AMPS/CDMA Tx	836.5	35	4.18	1.8
CDMA IF	85.38	1.26	1.48	12.00
CDMA IF	130.38	1.23	0.94	7.50
CDMA IF	183.6	1.26	0.69	10.50
CDMA IF	210.38	1.26	0.60	9.00
CDMA IF	220.38	1.26	0.57	8.00
CDMA Base Tcvr Subsys(BTS)	70	9.4	13.43	28.00
CDMA BTS	150	1.18	0.79	25.00
Broadband access	1086	10	0.92	5.00
Broadband access	499.25	1	0.20	9.00
Broadband access	479.75	23	4.79	13.00
Broadband access	333	0.654	0.20	9.00
Cable TV tuner	1220	8	0.66	4.2
DCS	1842.5	75	4.07	3.6
EGSM	942.5	35	3.71	2.6
GSM	947.5	25	2.64	3
GSM	902.5	25	2.77	3
GSM IF	400	0.49	0.12	6.5
GSM BTS	71	0.16	0.23	9
GSM BTS	87	0.4	0.46	7
GSM BTS	170.6	0.3	0.18	9
PCS	1960	60	3.06	2.4
PCS	1880	60	3.19	2
W-CDMA	1960	60	3.06	2.1
W-CDMA IF	190	5.5	2.89	8
W-CDMA IF	380	5	1.32	9
W-LAN	900	25	2.78	6
Satellite IF	160	2.5	1.56	25.1
Satellite IF	160	4	2.50	24.8
GPS	1575.42	50	3.17	1.4
Wireless data	2441.75	83.5	3.42	5
Wireless data	770	17	2.21	7
Wireless data	570	17	2.98	17.2
Wireless data	374	17	4.55	10.5
Wireless data	240	1.3	0.54	11

Table 1.	Commerciall	y Available	e SAW Filt	ters for Te	lecommunication
----------	-------------	-------------	------------	-------------	-----------------

an LNA. Thus the two channels are multiplexed by frequency division.

The filters used in these applications require low insertion loss and large stopband rejection, as well as small transition band. For these reasons, SAW filters are excellent candidate devices. Figure 30 shows a general schematic for a ladder-type filter. The parallel and series impedances, Z_p and Z_s , are created using SAW resonator (Fig. 24) structures. Additional inductance due to bond wires are also shown as part of the ladder network as they contribute to the performance at gigahertz-range operating frequencies.

5.2. Cellular Phone Transceiver Modules

Both analog (e.g., AMPS) and digital (e.g., GSM) cellular phone technologies employ several SAW filters and oscillators. As an example, Fig. 36 depicts an

AMPS cellular phone handset. AMPS employs frequencydivision multiple access, transmitting at 824–859 MHz and receiving at 869–894 MHz. The transceiver system employs six SAW devices. The antenna duplexer for the 800-MHz transmit and receive carriers is built from two SAW filters labeled Rx1 and Tx1. Additional carrier filtering is provided by SAW Rx2 and Tx2. The IF stage and VCO and PLL synthesizer also utilize SAW devices.

The SAW devices used in the receiver, Rx1 and Rx2, are typically of the LSAW resonator type. They are required to suppress harmonics, reject image frequency noise, and suppress switching noise from the power supplies and power amplifiers. The transmit filter, Tx1 and Tx2, is also typically an LSAW device, especially as it is required to handle up to 1 W of transmit power. The IF stage SAW filter is required to be very selective (30-kHz spacing) and very stable. Therefore this filter is often a waveguidecoupled resonator built on ST-cut quartz. The VCO and



Figure 36. Simplified block diagram of AMPS cellular phone handset depicting the use of up to six SAW devices in a 800-MHz dual heterodyne transceiver circuit [2].

PLL SAW devices are often dual-mode resonator filters or wideband delay lines. These same filter attributes are required in digital phone systems as well, the operating frequencies are simply altered accordingly.

5.3. Nyquist Filters for Digital Radio Link

Digital communication systems require a Nyquist filter for bandlimiting and to prevent intersymbol interference (ISI) [24]. The details of Nyquist filters are developed in the many texts on the subject of digital communications. Stated briefly, the Nyquist criterion for a zero ISI filter requires that frequency-domain sum of the entire filter spectrum be a constant. Or to put it another way, that the digital pulse spectrum is band-limited. One such filter that meets this criterion is the raised-cosine (RC) filter. Digital microwave radios often employ raised cosine filters in both the transmit and receive sides. In this way, the received matched filter response is raise-cosinesquared, which also can be shown to satisfy the Nyquist criterion. SAW filters are used to implement raised cosine and similar pulseshaping Nyquist filters for the North American common microwave carrier bands (4, 6, 8, and 11 GHz) [2].

5.4. Convolvers as Correlators for Spread-Spectrum Receivers

SAW convolvers are used in spread-spectrum communication links because of their small size, large processing gain, and broad bandwidth [2]. Typically a SAW convolver is implemented at the IF rather than RF stage. Actually, the convolver is used to perform autocorrelation of pseudonoise spreading codes. One port of the convolver (see Figs. 25–28) receives the IF coded sequences, while the second port receives the time-reversed reference code. The autocorrelated output is obtained at the third port. The interaction area, such as the semiconductor strip in a separated media structure, must be long enough to hold an entire bit sequence. One of the inputs, most likely the reference sequence, must be strong enough to generate the nonlinear action within the SAW substrate, or separated strip.

SAW convolver operating frequencies include the 900-MHz band, the 2-GHz spread-spectrum band, and the Japanese license-free spread-spectrum band below 322 MHz.

5.5. Clock Recovery Circuit for Optical Communications Link

SAW oscillators and filters are used in clock recovery circuits for optical communication links. The key advantages of SAW again lie in their small size, stability, and low jitter performance. SAW filters are used at center frequencies that correspond to the ATM/SONET/SDH clock frequencies. Example operating frequencies include 155.52, 622.08, and 2488.32 MHz. SAW oscillators are often used in PLL circuits as local oscillator references and VCO references.

These are some of the applications of SAW devices. There are others, including Global Positioning System (GPS) receivers, pagers, and ID tags—almost every telecommunication application. Additional applications are available in the literature [4,5,18,19].

BIOGRAPHY

Robert J. Filkins received the B.S. and M.S. degrees in electrical engineering in 1990 and 1997, respectively, from Rensselaer Polytechnic Institute, Troy, New York. He joined General Electric in 1993 as an Electronics Engineer, developing low-noise electronics systems for ultrasonic inspection of turbine components. Since 1995, he has been with the General Electric Global Research Laboratory, working on low-noise amplifier systems for ultrasonic, eddy-current, and laser ultrasonic inspection of components, as well as wireless and photonic communication components. Mr. Filkins currently holds six patents in the areas of electronics and nondestructive inspection systems. His areas of interest include optoelectronic devices and materials, laser generation of SAW, and microwave design. He is currently working toward a Ph.D. degree at Rensselaer Polytechnic Institute.

BIBLIOGRAPHY

- P. K. Das, Optical Signal Processing Fundamentals, Springer-Verlag, New York, 1991.
- 2. B. A. Auld, *Acoustic Fields and Waves in Solids*, Vol. II, 2nd ed., Krieger Publishing, 1990.
- 3. W. P. Mason, *Physical Acoustics*, Vol. I, Part A, Academic Press, New York, 1964.
- 4. C. K. Campbell, Surface Acoustic Wave Devices for Mobile and Wireless Communications, Academic Press, New York, 1998.
- E. A. Ash et al., in A. A. Oliner, ed., Acoustic Surface Waves, Topics in Applied Physics Vol. 24, Springer-Verlag, New York, 1978.
- 6. J. deKlerk, Physics Today 25: 32-39 (Nov. 1972).
- 7. L. B. Milstein and P. K. Das, *IEEE Commun. Mag.* **17**: 25–33 (1979).
- F. G. Marshall, C. O. Newton, and E. G. S. Paige, *IEEE Trans.* SU-20: 124–134 (1973).
- 9. W. R. Smith et al., IEEE Trans. MTT-17: 865-873 (1969).
- C. S. Hartmann, W. J. Jones, and H. Vollers, *IEEE Trans.* SU-19: 378–384 (1972).
- R. M. Hays and C. S. Hartmann, Proc. IEEE 64: 652–671 (1976).
- A. Chatterjee, P. K. Das, and L. B. Milstein, *IEEE Trans.* SU-32: 745–759 (1985).
- T. Martin, Low sidelobe IMCON pulse compression, Ultrasonic Symp. Proc. (J. deKlerk, ed.), IEEE, New York, 1976.
- A. Budak, Passive and Active Network Analysis, Houghton-Miflin, Boston, 1973, Chap. 4.
- S. Lehtonen et al., SAW Impedance Element Filters towards 5 GHz, *IEEE Ultrasonics Symp.*, 1998, pp. 369–372.
- T. A. Fjeldy and C. Ruppel, eds., Advances in Surface Acoustic Wave Technology, Systems, and Applications, Vol. 1, World Scientific, 2000.
- Y. Fujita et al., Low loss and high rejection RF SAW filter using improved design techniques, *IEEE Ultrasonics Symp.*, 1998, pp. 399-402.
- G. Kino, Acoustic Waves: Devices, Imaging, and Analog Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1987.
- D. P. Morgan, Surface Wave Devices for Signal Processing, Elsevier, New York, 1985.
- 20. R. C. Williamson and H. I. Smith, *IEEE Trans.* SU-20: 113–123 (1973).
- P. K. Das and W. C. Wang, Ultrasonics Symp. Proc., IEEE, New York, 1972, p. 316.
- 22. E. D. Palik, Handbook of Optical Constants of Solids, Academic Press, New York, 1997.
- A. J. Slobodnik, E. D. Conway, and R. T. Delmonico, Microwave Acoustics Handbook, Vol. IA, Surface Wave Velocities, AFCRL-TR-73-0597, Air Force Cambridge Research Labs, Cambridge, MA.
- J. G. Proakis, *Digital Communications*, 2nd ed., McGraw Hill, New York, 1989, pp. 532–536.

SURVIVABLE OPTICAL INTERNET

HUSSEIN T. MOUFTAH PIN-HAN HO Queen's University at Kingston Kingston, Ontario, Canada

1. INTRODUCTION

The Internet has revolutionized the computer and communications world like nothing before, which has changed the lifestyles of human beings by providing at once a worldwide broadcasting capability, a mechanism for information dissemination, and a medium for collaboration and interaction between individuals and their computers without regard for geographic location. As the importance of the Internet is overwhelming, the strategies of constructing the infrastructure are becoming more critical. The design of the Internet architecture with a suite of control and management strategies, by which the most efficient, scalable, and robust deployment can be achieved, has been a focus of researches in industry and academia since the late 1980s.

The Internet is a core network, or a wide-area network (WAN), which is usually across countries, or continents, for the purpose of interconnecting hundreds or even thousands of small-sized networks, such as metropolitanarea networks (MANs) and local area networks (LANs). The small-sized networks are also called *access networks*, which upload or download data flows to or from the core network through a traffic grooming mechanism.

The proposal of constructing networks on fiberoptics has come up with the progress in the photonic electronics since the 1960s or so, however, most of which were focused on the access networks before 1990, such as the fiber distributed data interface (FDDI) and synchronous optical network (SONET) ring networks. With the advances in the optical technologies, most notably dense wavelengthdivision-multiplexed (DWDM) transmission, the amount of raw bandwidth available on fiberoptic links has increased by several orders of magnitude, in which a single fiber cut may influence a huge amount of data traffic in transmission. As a result, the idea of using optical fibers to build the infrastructure of the Intern along with survivability issues emerged in the early 1990s. The adoption of a survivable optical Internet is for the purpose of reducing the network cost and enabling versatile multimedia applications by a simplification of administration efforts and a guarantee of service continuity (or system reliability) during any network failure.

This article presents the state-of-the-art progress of the survivable *optical Internet*, and provides thorough overviews on a variety of aspects of technologies and proposals for the protection and restoration mechanisms. To provide enough background knowledge, the following subsections describe the evolution of data communication networks, as well as the definition of *survivability* as the introduction of this article.

2462 SURVIVABLE OPTICAL INTERNET

1.1. Evolution of Data Communication Networks

Since the late 1980s, router-based cores have been largely adopted, where traffic engineering (TE) (i.e., the manipulation of traffic flow to improve network performance) was achieved by simply manipulating routing metrics and link states of interior gateway protocol (IGP), such as open shortest path first (OSPF) and intermediate system—intermediate system (IS-IS). As the increase in Internet traffic of new applications such as multimedia services, a number of limitations in terms of bandwidth provisioning and TE manipulation came up:

- 1. Software-based routers had the potential of becoming traffic bottlenecks under heavy load. Hardwarebased IP switching apparatus (which is defined below), however, are vendor-specific and expensive.
- 2. Static metric manipulation was not scalable and usually took a trial-and-error approach rather than a scientific solution to an increasingly complex problem. Even after the addition of TE extensions with dynamic metrics such as maximum reservable bandwidth of links to the IGPs, the improvements remain very limited [1,2].

To overcome the problems of bandwidth limitation and TE requirements, application-specific integrated circuit (ASIC)-based switching devices had been developed since the early 1990s. As prices of such devices were getting cheaper, IP over asynchronous transfer mode (ATM) was largely adopted and provided benefits for the Internet service providers (ISPs) in the aspects of the high-speed interfaces, deterministic performance, and TE capability by manipulating permanent virtual circuitry (PVC) in the ATM core networks.

While the IP over ATM core networks mentioned above solved the problems in the IP routing cores for the ISPs, however, the expense paid for the complexity of the overlay model of the IP over ATM networks has motivated the development of various multilayer switching technologies, such as Toshiba's IP switching [3] and Cisco's tag switching [4].

Multiprotocol label switching (MPLS) is the latest step in the evolution of multilayer switching in the Internet. It is an Internet Engineering Task Force (IETF) standardsbased approach built on the efforts of the various proprietary multilayer switching solutions. MPLS is composed of two distinct functional components: a control component and a forwarding component. By completely separating the control component from the forwarding component, each component can be independently developed and modified. The only requirement is that the control component continues to communicate with the forwarding component by managing the packet forwarding table. When packets arrive, the forwarding component searches the forwarding table maintained by the control component to make a routing decision for each packet. Specifically, the forwarding component examines information contained in the packet's header, searches the forwarding table for a match, and directs the packet from the input interface to the output interface across the system's switching fabric.

The forwarding component of MPLS is based on a label-swapping forwarding algorithm. This is the same algorithm used to forward data in ATM and frame relay switches [1,2,5]. An MPLS packet has a 32-bit header carried in front of the packet to identify a forwarding equivalence class (FEC) [5]. A label is analogous to a connection identifier, which encodes information from the network layer header, and maps traffic to a specific FEC. An FEC is a set of packets that are forwarded over the same path through a network even if their ultimate destinations are different.

1.2. Survivability Issues

A network is considered to be survivable if it can maintain service continuity to the end users during the occurrence of any failure by preplanned or real-time mechanisms of protection and restoration. Network survivability has become a critical issue as the prevalence of DWDM, by which a single fiber cut may interrupt huge amount of bandwidth in transmission. In general, Internet backbone networks are overbuilt in comparison to the average traffic volumes, in order to support fluctuations in traffic levels, and to stay ahead of traffic growth rates. With the underutilized capacity in networks, the most widely recognized strategy to perform protection service is to find protection resources that are physically disjoint (or diversely routed) from the working paths, over which the data flow could be switched to the protection paths during any failure of network elements along the working paths.

This article is organized as follows. Section 2 presents a framework of interest in this article for performing protection and restoration services in the *optical Internet*, and includes some background knowledge and important terminology for the subsequent discussions. Section 3 describes several classic approaches of achieving *survivability* with implementation issues. Section 4 investigates into the strategies newly reported in the literature. Section 5 summarizes this article.

2. FRAMEWORK FOR ACHIEVING SURVIVABILITY

This section focuses on the framework of protection and restoration for the *optical Internet*. Some of the most important background knowledge and concepts in this area are presented.

2.1. Background

Although the use of DWDM technology enables a fiber to accommodate a tremendous amount of data; it may also risk a serious data loss when a fault occurs (e.g., a fiber cut or a node failure), which could downgrade the service to the customers to the worst extent. To improve the *survivability*, the ISPs are required to equip the networks with protection and restoration schemes that can provide end-to-end guaranteed services to their customers according to the service-level agreements (SLAs).

Faults can be divided into four categories: path failure (PF), path degraded (PD), link failure (LF), and link degraded (LD) [6]. PD and LD are cases of loss of signal (LoS) in which the quality of the optical flow is

unacceptable by the terminating nodes of the *lightpath*. To cope with this kind of failure, Hahm et al. [7] suggested that a predetermined end-to-end path that is physically disjoint from the working path is desired, since a fault localization cannot be conducted with respect to LoS along the intermediate optical network elements in an all-optical network. Because the probability of occurrence of an LoS fault in the optical networks is reported to be as low as 1×10^{-7} [8], it rarely needs to be considered. Furthermore, an LoS fault is generally caused by a failed transmitter or wavelength converter, which could be overcome by switching the traffic flowing on the impaired channel to the spare wavelength(s) along the same physical conduit. Since this article focuses on *shared protection* schemes in which protection paths are physically disjoint from working paths, the protection and restoration of an LoS fault on a channel is not included in this paper. Gadiraju and Mouftah [9] have further discussed and analyzed this topic.

In the PF and LF cases, the continuity of a link or a path is damaged (e.g., a fiber cut). This kind of failure can be detected by a loss of light (LoL) detection performed at each network element so that fault localization [7] can be easily performed. The restoration mechanisms may have the protection path traversing the healthy NEs along the original working path as much as possible while circumventing from the failed NEs to save the restoration time and improve network resources. In this article, all nodes are assumed to be capable of detecting an LoL fault in the optical layer. Optical detectors residing in the optical amplifiers at the output ports of a node monitor the power levels in all outgoing fibers. An alarm mechanism is performed at the underlying optical layer to inform the upper control layer of a failure once a power-level abnormality is detected.

In this article, protection and restoration schemes are discussed and evaluated with the following criteria: scalability, dynamicity, class of service, capacity efficiency, and restoration speed. The scalability of a scheme is important in a sense that networks are required to be scalable to any expansion in the number of nodes or edges with the same computing power in the control center or in each node. The scalability ensures that the scheme can be applied to large networks in size and capacity. The dynamicity of a scheme determines whether the networks can deal with dynamic traffic that arrives at the networks one after the other without any prior knowledge, which makes the optimization of spare capacity a challenge. The networks with class of service can provide the end users with protection services with wider spectrum and finer granularity, and are the basis on which the ISPs charge their customers. The *capacity* efficiency is concerned with the cost of networks. To be capacity-efficient, the schemes should make the most use of every piece of spare capacity through optimization or heuristics. Restoration speed describes how fast a failure can be recovered after its occurrence. However, restoration speed, capacity efficiency, and dynamicity are tradeoffs in the design spectrum of network protection and restoration schemes most of the time.

Protection can be defined as the efforts made before any failure occurs, including the failure detection and localization. *Restoration*, on the other hand, is the reaction of the protocol and network apparatus toward the failure, including any signaling mechanisms and network reconfiguration used to recover from the failure. A *lightpath* is defined as a data path in an optical network that may traverse one or several nodes. A *working path* (or primary path) is defined as a *lightpath* that is selected for transmitting data during normal operation. A *protection path* (or spare path, secondary path) is the path used to protect a specific segment of working path(s).

Various types of protection and restoration mechanisms have been reported, and can be categorized as *dedicated* and shared protection in terms of whether the spare capacity can be used by more than one working paths. The *dedicated protection* can be either 1 + 1 or 1:1. The 1+1 protection is characterized by having two disjoint paths transmitting the data flow simultaneously between sender and destination nodes, thus an ultrafast restoration speed is achieved. As a failure occurs, the receiver node does not have to switch over the traffic flow along the working path for keeping the service continuity. The 1:1 protection, on the other hand, has a *dedicated protection* path for the working path without passing data traffic along the protection path. In other words, the network resources along the protection path is not configured, and can be used for the other traffic with a lower priority during normal operation (this kind of low-priority traffic is also called "best-effort" traffic). Once a failure occurs, the best-effort traffic has to yield the right of way, so that the traffic in the affected working path is switched over to the corresponding protection path.

The shared protection has versatile types of design originality, which can be categorized as path-based, linkbased, and short leap shared protection (SLSP) according to the location where fault localization is performed. In the case of a number of working paths sharing a protection path, shared protection can be either 1:N or M:N. In the 1:N case, a protection path is shared by N physically disjoint working paths. In the M:N case, M protection paths are shared by N working paths. The working paths sharing the same protection path are required to be physically disjoint because a protection path can afford a switchover of a single working path only at the same moment. This is also called a shared risk link group (SRLG) constraint, which will be discussed in later sections.

Opposite to the investigation into the relationship between paths in the networks, proposals were made to preplan or preconfigure spare capacity at the network planning stage according to the working capacity along each span (or edge). The planning-type schemes are used in networks with static traffic that is rarely changed as time goes by, and usually requires a time-consuming optimization process (e.g., integer linear programming or some flow theories).

2.2. Shared Risk Link Group (SRLG) Constraint

The *shared risk link group* (SRLG) [10] is defined as a group of working *lightpaths* that run the same risk of

service interruption by a single failure, which has the following characteristics:

$$\operatorname{SRLG}(P_{M,u}^m) = \bigcup_{M \cap Q \neq \varnothing} \bigcup_{k \in K} \{P_{Q,k}^m\},$$
$$P_1 \in \operatorname{SRLG}(P_2) \Leftrightarrow P_2 \in \operatorname{SRLG}(P_1)$$

where u and k are wavelength planes (which could be the same in the case of a multifiber system), and M and Qare two sets of optical network elements (ONEs) traversed by the two lightpaths $P_{M,u}^m$ and $P_{Q,k}^m$. The SRLG of the lightpath $P_{M,u}^m$, SRLG ($P_{M,u}^m$), is the union of all lightpaths (which belong to the set of all wavelength planes, K) that run the same risk of a single failure with $P_{M,u}^m$. In other words, they traverse at least one common ONE.

SRLG is a dynamic link state that needs to be updated whenever a *lightpath* is modified (e.g., a teardown or a buildup). SRLG is hierarchical [10], and is not limited to physical components (but also protocols, etc). The SRLGs existing in the network can be derived by the following pseudocode:

```
For n = 1 to N do
   Derive all lightpaths traversing the nth ONE
     and put them into temp;
   covered = false;
   If temp \neq \emptyset
      For l = 1 to N do
         If temp \supset S_l
                        //S_l can not be a root element
             S_l \leftarrow \emptyset
         End if
         If temp \subset S_l and S_l \neq \emptyset
             covered = true; // the nth ONE is not a root element
              Break:
         End if
      End for
      If covered = false
          S_n \leftarrow temp
       End if
   End if
End for
```

Here, S_l and *temp* are lightpath sets. S_l represents the SRLG with all its lightpaths overlapping at the *l*th ONE. A root element is where the corresponding SRLG is defined, in which all the lightpaths traverse through the ONE. To update the SRLG information on-line, the processing of the algorithm should be finished before the next event arrives.

The SRLG constraint is a stipulation on the selection of protection resources with which a ONE cannot be reserved for protection by two or more working lightpaths if they belong to the same SRLG. The purpose of following the SRLG constraint is to guarantee the 100% survivability for a single failure occurring to any ONE in the network. When *shared protection* is adopted, in the case that a working and protection paths can be on different wavelength planes, the SRLG constraint impairs the performance more than the case where working and protection paths have to be on the same wavelength plane. It is obvious that the former situation outperforms the latter because of a more flexible utilization of wavelength channels; however, the expense paid for the better performance is that more tunable transceivers and optical amplifiers are required in a node. Note that in the latter, the SLRG constraint exists only if the system is a multifiber system, in which two or more wavelength channels on the same wavelength plane could be in the same SRLG. The lightpaths that take the same wavelength plane may suffer from the SRLG constraint since they cannot share the same protection resources (e.g., a wavelength channel) if they belong to the same SRLG.

As shown in Fig. 1a, working paths P1 and P2 are in the same SRLG with each other since the two working paths have a physically overlapped span F-G. To restore P1 and P2 at the same time once a failure occurred on span F-G, the number of spare links prepared for P1 and P2 should be the summation of bandwidth of P1 and P2 along the spans A-K-H-I-J-D-G. On the other hand, for the two path segments P1 (S-E-F) and P2 (A-B-F-G) in which there is no overlapped span, as shown in Fig. 1b, they are not in the same SRLG. The spare capacity along the spans A-K-H-I-J-D for P1 and P2 can be the maximum of the two paths.

3. CONVENTIONAL TECHNIQUES

This section introduces some classical techniques for protection and restoration, which are frameworks proposed by both industry and academia. The topics included are *SONET self-healing ring*, path-based, link-based *shared protection* schemes, and *short leap shared protection* (SLSP).

The first scheme, SONET self-healing ring, is to set up a network in a ring architecture, or in a concatenation of rings, with the facilitation of standard signaling protocols. SONET self-healing rings are characterized by stringent recovery time scales from a failure and a recovery service time of 50 ms. The restoration process performed within the recovery time includes detection, switching time, ring propagation delays, and resynchronization, which are derived from a frame synchronization at the lowest frame speed, (DS1, 1.5 Mb/ps) [8]. On the other hand, the expense paid for the restoration speed is the capacity efficiency, in which more than 100% (and up to 300%) of capacity redundancy is required.

The second and third schemes are designed for a mesh network, which is characterized by high-*capacity efficiency*, expensive optical switching components, and



Figure 1. Since there is an overlapped span F-G for P1 and P2 in (**a**), the spare capacity along A-K-H-I-J-D-G has to be the summation of the two working paths. In (**b**), since there is no overlapped span and node, the spare capacity along A-K-H-I-J-D can be shared by P1 and P2.

relatively slow restoration services compared with the SONET self-healing ring.

3.1. SONET Self-Healing

This subsection introduces the *SONET self-healing ring*, which has been an industry standard. The content in this subsection is adopted mainly from the *Internet Draft* [11].

Two types of SONET self-healing rings are widely used: (1) a *unidirectional path-switched ring* (UPSR), which consists of two unidirectional counter-propagating fiber rings, referred to as *basic rings*; and (2) a two-fiber (or fourfiber) bidirectional line-switched ring (BLSR/2 or BLSR/4), which consists of two unidirectional counter-propagating basic rings as well.

In UPSR, the basic concept is to design for channel level protection in two-fiber rings. UPSR rings dedicate one fiber for working time-division multiplexing channels (TDM time slots) and the other for corresponding protection channels (counterpropagating directions). The Traffic is permanently sent along both fibers to exert a 1+1 protection. Different rings are connected via bridges. To satisfy a bidirection connection, all the resources along working and protection fibers are consumed, as a result, the throughput is restricted to that of a single fiber.

Clearly, UPSR rings represent simpler designs and do not require any notification or switchover signaling mechanisms between ring nodes, namely, receiver nodes perform channel switchovers. As such, they are resourceinefficient since they do not reuse fiber capacity (both spatially, and between working and protection paths). Moreover, span (i.e., fiber) protection is undefined for UPSR rings, and such rings are typically most efficient in access rings where traffic patterns are concentrated around collector hubs.

As for the BLSR, it is designed to protect at the line (i.e., fiber) level, and there are two possible variants, namely two-fiber $(BLSR\!/\!2)$ and four-fiber $(BLSR\!/\!4)$ rings. The BLSR/2 concept is designed to overcome the spatial reuse limitations associated with two-fiber UPSR rings and provides only path (i.e., line) protection. Specifically, the BLSR/2 scheme divides the capacity timeslots within each fiber evenly between working and protection channels with the same direction (and having working channels on a given fiber protected by protection channels on the other fiber). Therefore bidirectional connections between nodes will now traverse the same intermediate nodes but on differing fibers. This allows for sharing loads away from saturated spans and increases the level of spatial reuse (sharing), a major advantage over two-fiber UPSR rings. Protection slots for working channels are preassigned on the basis of a fixed odd/even numbering scheme, and in case of a fiber cut, all affected time slots are looped back in the opposite direction of the ring.

This is commonly termed "loopback" line/span protection and avoids any per-channel processing. However, loopback protection increases the distance and transmission delay of the restored channels (nearly doubling pathlengths in the worst case). More importantly, since BLSR rings perform line switching at the switching nodes (i.e., adjacent to the fault), more complex active signaling functionality is required. Further bandwidth utilization improvements can also be made here by allowing lower-priority traffic to traverse on idle protection spans. Four-fiber BLSR rings extend on the BLSR/2 concepts by providing added span switching capabilities. In BLSR/4 rings, two fibers are used for working traffic and two for protection traffic (counterpropagating pairs, one in each direction). Again, working traffic can be carried in both directions (clockwise, counterclockwise), and this minimizes spatial resource utilization for bi-directional connection setups. Line protection is used when both working and protection fibers are cut, looping traffic around the long-side path. If, however, only the working fiber is cut, less disruptive switching can be performed at the fiber level. Here, all failed channels are switched to the corresponding protection fiber going in the same direction (and lowerpriority channels preempted).

Obviously, the BLSR/4 ring capacity is twice that of the BLSR/2 ring, and the four-fiber variant can handle more failures. In addition, it should be noted that both two- and four-fiber rings provide node failure recovery for passthrough traffic. Essentially, all channels on all fibers traversing the failed node are line-switched away from the failed node. BLSR rings, unlike UPSR rings, require a protection signaling mechanism. Since protection channels can be shared, each node must have a global state, and this requires state signaling over both spans (directions) of the ring. This is achieved by an automatic protection switching (APS) protocol, or also commonly termed SONET APS. This protocol uses a 4-bit node identifier and hence allows only up to 16 nodes per ring. Additional bits are designated to identify the type of function requested (e.g., bidirectional or unidirectional switching) and the fault condition (i.e., channel state). Control nodes performing the switchover functions utilize framepersistency checks to avoid premature actions and discard any invalid message codes.

3.2. Path-Based Shared Protection

For a path-based *shared protection*, as shown in Fig. 2, the first hop node [6] of the working path w1 computes the protection path p1, which has to be diversely routed from the working path according to the SRLG information. If a fault occurs on the working path, whether it is an LoL or an LoS, the terminating node (N1) in its control plane realizes the fault and sends a notification indicator



Figure 2. Ordinary path-based 1:N protection.

signal (NIS) [6] to the first hop node of the path to activate a switchover. Then, the first hop node immediately sends a wakeup packet to activate the configuration of the nodes along the protection path and then switch over the whole traffic on the working path to the protection path.

One of the most important merits of the pathbased protection scheme is that it handles LoS and LoL in a single move with relatively lower expense of network resources that need to be reserved. In addition, restoration becomes simpler in terms of signaling algorithm complexity since only the terminating node of the path needs to respond to the fault. On the other hand, it incurs the following difficulties and problems: (1) the complexity of calculation for the diverse protection route grows fast with the increasing number of nodes in the domain, and (2) the protection resources cannot be shared by any other working path that violates the SRLG constraint with the protected working path. For example, in Fig. 2, p1, the protection path for w1, cannot share any of its resources to protect w2 because w2 shares the same link group with w1 only in 1 out of the 17 links.

3.3. Link-Based Shared Protection

Link-based protection was originally devised for the networks with a ring-based architecture such as SONET, in which the effects of network planning play an important role in performance. The scalability and reconfigurability have been criticized [12,13] in that a global reconstruction may be realized as a result of a small localized change, which limits the network's scale. The migration of linkbased protection from ring-based networks to mesh networks needs more network planning efforts and heuristics, which has been explored intensely [14-16]. In general, the link-based protection in mesh networks is defined as a protection mechanism that performs a fault localization during the occurrence of a failure, and restores the interrupted services by circumventing the traffic from a failed link or node at the upstream neighbor node and merging back to the original working path at the downstream neighbor node. In other words, to protect both the downstream neighbor link failure and node failure, two merge nodes have to be arranged for every node along a working path: one for the upstream neighbor link and one for the upstream neighbor node. As shown in Fig. 3, node



Figure 3. Link-based protection in a mesh WDM network. The node localizing an upstream fault behaves as a PML *path merge LSR* (*label switch router*) (PML) [6], which only needs to notify its upstream neighbor that behaves as a *path switch LSR* (PSL) [6] before traffic can be switched over to the protection links. The time for transmitting NIS is totally saved in this local restoration mechanism.

A has B and E as merge nodes for w1 and has B and C as merge nodes for w2.

With the associated signaling mechanisms, link-based protection provides a faster restoration due to the fault localization and a better throughput due, in turn, to the relaxation of the SRLG constraint. However, the large amount of protection resources consumed by this scheme may impair the performance and leave room for improvement. Because of its design for the ring-based architecture and its goal for protecting links between nodes, an end-to-end protection service is hard to perform by a pure link-based protection. In addition, it is debatable whether the downstream neighbor node and link are required to have separate protection circles. In Section 3.4, a new definition of link-based protection for achieving an end-to-end protection service in mesh networks will be introduced.

3.4. Short Leap Shared Protection (SLSP)

SLSP [18,19] is an end-to-end service-guaranteed shared protection scheme, which is an enhancement of the linkand path-based shared protection, for providing finer service granularities and more network throughput. The main idea of SLSP is to subdivide a working path into several fixed-size and overlapped segments, each of which is assigned by the first hop node a protection domain ID (PDID) after the working path is selected, as shown in Fig. 4. The diameter of a protection domain (or *P* domain) is defined as the hop count of the shortest path between the PSL and PML in the P domain. The definition of SLSP generalizes the shared protection schemes, in which the link- and path-based shared protection can be categorized as two extreme cases of SLSP with domain diameters of 2 and H, respectively, where H is the hop count of the working path.

The protection path can be calculated for each *P* domain either by the first hop node alone, or distributed to the PSL in each *P* domain, depending on how the SRLG information is configured in each node and how heavy a workload the first hop node can afford at that time. Figure 4 illustrates how a path under SLSP is configured and recovered when a fault occurs. Node *A* is the first hop node, and node *N* is the last hop node, which could respectively be the source node and the destination node of this path. The first *P* domain (PDID = 1) starts at node *A* and ends at node *F*. The second *P* domain (PDID = 2) is from node *E* to node *J*, and the third is from node *I* to node *N*. In this case, (*A*,*F*), (*E*,*J*), and (*I*,*N*) are the corresponding PSL-PML pairs for each *P* domain.

Since each P domain is overlapped with its neighboring P domains by a link and two nodes, a single failure on any link or node along the path can be handled by at least one P domain. If a fault occurs on the working path, the Link Management Protocol (LMP) helps localize the fault, and the PSL of the P domain in which the fault is located will be notified to activate a traffic switchover. For example, a fault on link 4 or node E is localized by node D. A fault on link 5 or node F is localized by node E. In the former case, node D sends a *notification indicator signal* (NIS) to notify node A that a fault occurred in their P domains. In the later case, node E is itself a PSL. In each



Figure 4. SLSP protection scheme divides the working path into several overlapped P domains. Nodes A, E, I are the PSLs, and nodes F, J, N are the PMLs.

case, the PSL (i.e., A or E) immediately sends a wake-up packet to activate the configuration of each node along the corresponding protection path, and then the traffic can be switched over to the protection path. A tell-andgo (TAG) [17] strategy can be adopted at this moment so that the PSL (i.e., node A or E) may switch the traffic to the protection path before an acknowledgment packet is received from the PML (i.e., node F or I). At completion of the switchover, the information associated with this rearrangement has to be disseminated to all the other nodes so that the best-effort traffic will not be arranged to use these resources. Under the single-failure assumption, it is impossible that more than one working path will switch their traffic to protection paths at the same time. However, for the environment where multiple failures are considered, a working path has to possess two or more sets of partially or totally disjoint protection paths to prevent the possibility that its protection resources are busy while it needs them. When the fault on the working path is fixed and a switchback to the original working path is required, a notification for releasing the protection resources has to be sent by the PSL to the first hop node right after the traffic is switched. With this, the protection resources can be reported as "free" again to all the other nodes at the next OSPF dissemination.

To implement the protection information dissemination, the association of the protection resources in each P domain with corresponding working path segments (PDID) has to be included in its forwarding adjacency. The other working paths must know this association before they can reserve any piece of protection resources for a protection purpose. An example is shown in Fig. 4, where w1 and w2 possess the same SRLG on link 8. However, w2 can share all the protection resources of w1 except those in the second P domain (PDID2). In addition, the signaling protocol, such as the resource reservation protocol (RSVP) or label distribution protocol (LDP), needs further extensions for the "path message" and "label request message" to carry object to assign PSLs and PMLs in the SLSP paradigm. The advantages of the SLSP framework over the ordinary path protection schemes are as follows:

- 1. The complexity of calculating a diverse route under the constraint of whole domain's SRLG information can be segmented and largely diminished to several *P* domains, in which the provisioning latency for dynamic path selection can be reduced.
- 2. Both the notification and the wakeup message may be performed only within a *P* domain; therefore, the restoration time is reduced according to the size of the *P* domain.
- 3. The protection service can be guaranteed more readily since the average/longest restoration time does not vary with the length of the whole path; instead, the average or largest size of the *P* domains will be the dominant factor, which can be an item with which the service providers bill their customers.
- 4. The computation complexity of protection paths is simplified by the segmentation of working paths. Section 4 introduces signaling issues in which the distributed allocation scheme can reduce the computation efforts of the first hop node that is usually a heavy-loaded border router.
- 5. The SRLG constraint is relaxed. SLSP divides a working lightpath into several segments, which results in two effects: an increase in the number of SRLGs in the network and a decrease in the size (or the number of lightpaths) of SRLGs. The former has little influence on the network performance while the later can improve the relaxation of SRLG constraint.

An example of advantage 5 (above) is illustrated in Fig. 5. In Fig. 5a, the two working lightpaths (A, G, H, I, F) and (A, G, J, K, L, M) are in the same SRLG if a path-based protection scheme is adopted, so that they cannot share any protection resources. On the other hand, with the adoption of SLSP, as shown in Fig. 5b, the same protection resources (G, R, S, T) can be shared by both of the working paths due



Figure 5. (a) Path-based protection; (b) SLSP with diameters of 2 and 3.

to the segmentation into two P domains. Note that the two lightpaths in Fig. 5 are in the same wavelength plane so that they are stipulated by the SRLG constraint.

Compared with the pure link protection approach, SLSP provides adaptability in compromising restoration time and protection resources required, with which the class of service can be achieved with more granularity. The Isps can put proper constraints on the path selection according to the service-level agreement with each of their customers. The constraining parameters for the selection of the protection path can be those related to the restoration time along the working path, such as the diameter of each P domain and the physical distance between each PSL-PML pair.

4. SPARE CAPACITY ALLOCATION

This section introduces the protection and restoration schemes, which are based on the efforts of optimization on the spare capacity allocation performed either at the network planning stage, or during the time interval between two network events (i.e., a connection setup or teardown) if each connection request arrives one after the other with a specific holding time. Since the optimization needs to know all the working paths in the network, it has to be processed in a static manner. In other words, any network changes, including a change of traffic, or an extension to the topology, or a change of network administrative policies and routing constraint, will require the optimization to be performed again. The static restoration schemes that will be presented in this section are ring cover/node cover [14,15], preconfigured cycle (or p-cycle) [20-22], protection cycle [23-25], survivable routing [26,27,29], and static SLSP [30].

4.1. Ring Cover/Node Cover

Algorithms were developed to cover all the link/node with least spare capacity in a shape of cycle, tree, or a mixture of both, which have been proved to be NP-hard. Once a fault occurs on any edge or node, a restoration process can be activated to switch over the traffic along the impaired edge onto the spare route. Since every edge/node is covered by the preplanned spare resource, a fault can be recovered any time and anywhere in the network [14,15].

The fatal drawback of *ring / node cover* is that the spare capacity along an edge has to be the maximum capacity of all edges in the network, so that all the possible failure events can be dealt with. This characteristic of *ring / node cover* has motivated the development of more capacity-efficient schemes based on the similar design originality.

Preconfigured cycle (or *p-cycle*) is one of the successful proposals as far as we can see, and will be presented in the next paragraphs.

4.2. Preconfigured Cycle

The preconfigured cycle (or p-cycle) is one of the most successful and well-developed strategies for performing spare capacity management at the network planning stage, which is an extension of ring cover. The structure of spare capacity deployed into networks is in a shape of a ring, which is where the "cycle" comes from. Different from ring cover, *p-cycle* addresses networks in which working and spare capacity can vary from link to link, so as to reduce the waste of segmentation of capacity by a ring-based architecture [20-22]. The spare capacity along each ring (or a "pattern" in the terminology of *p*-cycle) is preconfigured instead of being only preplanned, which means that the best-effort traffic can hardly utilize these resources. With knowledge of all the working capacity in networks, *p*-cycle formulates the optimization of the summation of spare capacity in each span into an integer linear programming (ILP) problem, in which different patterns are chosen from the prepared candidate cycles. The result of the optimization is the number of copies for each different cycle pattern, which can be 0 or any positive integer.

Inevitably, the deployment of ring-shaped patterns into mesh networks may result in excess spare links due to a uniform capacity along a ring. On the other hand, since *p-cycle* provides restoration only for the on-cycle and straddling failure, as shown in Fig. 6, intercycle resource sharing is not supported. This above observation motivates us to solve this problem from another point of view; more mesh-based characteristics are taken into account along the design spectrum, which has two ends on the pure ring-based and pure mesh-based design originality.

Because the *p*-cycle does not allow intercycle sharing, the relationship between working paths is not considered. The optimization problem is formulated as follows [20]:

Target:

Minimize

$$\sum_{j=1}^{S} c_j \cdot s_j$$

Subject to the following constraints:

$$s_j = \sum_{i=1}^{N_p} (x_{i,j}) \cdot n_i$$

 $y_{i,j}) \cdot n_i = w_j + r_j,$



Figure 6. (a) A pattern of *p*-cycle in the network. (b) An on-cycle failure is restored by using the rest of the pattern. (c, d) a straddling failure can be restored by either side of the pattern.

where j = 1, 2, ..., S, and

$$x_{i,j} = \begin{cases} 1 & \text{span } j \text{ has a link on pattern } i \\ 0 & \text{otherwise} \end{cases}$$
$$y_{i,j} = \begin{cases} 0 & \text{both nodes of span } j \text{ not on pattern } i \\ 1 & \text{span } j \text{ is on-cycle with the pattern } i \end{cases}$$

 $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ span j is straddling with the pattern i

For each of the parameters, n_j — number of copies of the pattern i; w_j , s_j — number of working and spare links on span j; r_j — spare links excess of those required for span j.

The process of solving the ILP for *p-cycle* is notorious with its NP completeness and a long computation latency, which can prevent the scheme from being used in networks with larger size or with a dynamic traffic pattern. Therefore, the dynamicity is traded for the optimized capacity efficiency and the swift restoration. Heuristics was provided to improve the computation complexity [22], in which different cost functions were adopted to address either capacity efficiency, traffic capture efficiency, or a mixture of both, at the expense of performance. After a ring is allocated, the covered working links are never considered in the subsequent iterations. The iteration ends when all the working links are covered. In considering both intracycle and intercycle protection, the cost function developed takes both *capacity efficiency* η_{capacity} and traffic capture efficiency η_{capture} into account.

4.3. Protection Cycles

Another important proposal for ring-based protection is protection cycles, which is inherent from node and ring cover, and is intended to overcome their drawbacks [23-25]. Algorithms were developed to cover each link with double cycles for both planar and nonplanar networks, so that *automatic protection switching* (APS) can be performed in an optical network with arbitrary mesh topologies. Protection cycles also suffer from the problem of scalability and requiring a global reconfiguration in response to a local network variation. In addition, the granularity of infrastructure is limited to four-fiber instead of two-fiber in case a wavelength partition is not considered. As a result, the system flexibility is reduced in the implementation of networks with smaller service granularity, such as middle-sized or metropolitan-area networks.

4.4. Survivable Routing

Survivable routing is defined as an approach to derive a better backup path by using the shortest path algorithm with a well-designed cost function and link metrics. Two types of *survivable routing* schemes are introduced in this section; (1) *successive survivable routing* (SSR) [26] and *asymmetrically weighted survivable routing* (AWSR).

In SSR, each traffic flow routes its working path first, then its backup path in the source node next so that the difference in importance between primary and secondary paths can be emphasized. This sequential derivation of the working and secondary paths is where the "successive" comes from. With SSR, protection paths are optimized according to each working path in the network instead of the working bandwidth along each edge, in order to facilitate the intercycle sharing of spare resource and to improve the *capacity efficiency*. For dealing with the relationship between working paths, the SRLG constraint has to be considered and included into the optimization process. Therefore, the optimization process can only be formulated into an integer programming (IP) problem with longer computation time consumed instead of an Integer Linear Programming problem, since the derivation of the SRLG relationship in networks can only be through a nonlinear operation.

In AWSR, on the other hand, working and protection paths are derived at the same stage with a weighting on the cost of a working path, namely, a cost function $\alpha \cdot C + CP$, where C and CP are costs of the working path and its corresponding protection path, respectively; and the parameter α is a weighting on the working path. The meaning of using the weighting parameter α is to distinguish the importance between the resources utilized by the working and protection paths, in which the derivation of the two paths is based on the same link metrics. A special case of $\alpha = 1$ can be solved by the SUURBALLE's algorithm [28] within a polynomial computation time. In case a dedicated protection (e.g., (1+1) is adopted, the best value of α could be as small as unity, since there is no difference to the consumption of network resources between a working path and a protection path. On the other hand, for a shared protection scheme (e.g., 1:N or M:N), the best value of α could be set large since the protection resources are shared by several working paths (or, in other words, the protection paths are less important than the working paths by α times).

In addition to the SUURBALLE's algorithm, the nodedisjoint diverse routing problem can be solved intuitively by the two-step algorithm [28], which first finds the shortest path, and then finds the shortest path in the same graph with the edges and nodes of the first path being erased. Although this method is straightforward and simple, it may fail to find any disjoint path pair after erasing the first path that isolates the source node from the destination node in the network. Besides, in some cases the algorithm cannot find the optimal path pair if there is another better disjoint path pair in which the working path is not the shortest one. To avoid these drawbacks, an enhancement for the two-step algorithm is necessary. The authors of Ref. 27 have conducted a study in which not only the shortest path is examined but also the loopless k shortest paths, where k = 1, 2, ..., K. The approach proposed in that paper [27] can also deal with the situation in which the working and protection paths are required to take different (or heterogeneous) link metrics, in order to further distinguish the characteristics of the two paths. On the other hand, another study [29] focused on the optimization or approximate optimization for the derivation of working and protection path pairs with uniform link metrics.

4.5. Static SLSP (S-SLSP)

S-SLSP is aimed at the task of an efficient spare capacity reallocation, which takes the *scalability* (or computation complexity) and network *dynamicity* into consideration [30]. In terms of resources sharing and *capacity efficiency*, the fundamental difference between S-SLSP and *p*-cycle or any other ring-based spare capacity allocation schemes is that the former investigates the relationship between lightpaths, so that all types of resource sharing are supported, which is potentially more capacity-efficient. As mentioned in Section 3.4, *P* domains are allocated in a cascaded manner along a working path, which facilitates not only link, but also node protection, so that an all-aspect restoration service can be provided. Because spare resources are preplanned instead of being preconfigured, better throughput can be achieved by launching best-effort traffic onto the spare capacity in normal operation.

There are two phases for the implementation of the S-SLSP algorithm:

- 1. Prepare all the cycles in the network up to a limited size.
- 2. Optimize the deployment of each cycle according to the existing working traffic in the network.

In step 1, the cycle listing algorithm can be seen in Grover et al. [22] and Mateti and Deo [31]. Step 2 can be simply formulated as an integer programming (IP) problem shown as follows.

Objective: Minimizing

$$\delta \cdot \sum_{j=0}^{S} c_j \cdot$$

 s_j

subject to the constraints

$$egin{aligned} s_j &= \max_i \{x_{i,j} \cdot ns_i\} + SR_j \ w_j &= \sum_{i=1}^{N_p} (x_{i,j}) \cdot nw_i \ n_i &= nw_i + ns_i \ W_j &\geq w_j + s_j \end{aligned}$$

where c_j and s_j are the cost and the spare capacity on the span j, respectively, and S is the number of spans in the network; n_i is the number of copies for the P domain i; SR_j is the extra spare links needed on span j for meeting the SRLG constraint; ns_i and nw_i are the numbers of spare and working links on span i; and W_i is the number of links on span i. The binary parameter $x_{i,j}$ is 1 if the span j has a link on domain i, and 0 otherwise. The binary parameter δ is 1 if the plan meets the requirement of SLSP, 0 otherwise.

Although the formulation above is feasible, the brutalattack-type optimization by examining the combinations of all cycles for each working path may yield terribly high computation complexity. S-SLSP reduces the computation complexity by interleaving the optimization process into several sequential sub-processes, in which an improvement from $O(2^N)$ to $O(m \cdot 2^n)$ can be made, where $N = n \cdot m$ is the number of candidate *P* domains for all working paths, m is the number of interleaved processes, and n is the number of candidate P domains in each subset of working paths. In other words, several subsets of working paths are optimized one after the other. It can be easily seen that the larger the number of working paths in a subset, the closer is the optimality that can be achieved, but at an expense of computation time.

Grouping of the existing working paths into several subsets is implemented according to the following two principles: (1) The working paths in a subset should be overlapped with each other as much as possible and (2) Since the candidate cycles can be up to a limited size according to the SLA of each working path, working paths with a similar requirement of restoration time are also grouped into the same subset. Working paths in a subset are processed with the integer programming solver together to determine the corresponding spare capacity. A framework of optimization on the performance can be developed with the size of each subset of working paths varied. Since every network event (i.e., a connection setup or teardown) will change the network capacity distribution, the spare capacity reallocation has to restart to follow the new link state. Therefore, a "successful" reallocation process is completed before the next network event arrives. Note that the computation time is exponentially increased with the volume of each subset. Since the increase in computation time will decrease the chance of completing the Integer programming task, as a consequence, the performance is impaired. On the other hand, the increase of the volume of each subset of working paths can improve the quality of optimization. Because of the two abovementional criteria, an optimized number of working paths in a subset can be derived to yield the best performance.

5. SUMMARY

The *optical internet* has changed the lifestyle of human beings by providing various applications, such as ecommerce, e-conferencing, Internet TV, Internet phone, and video on demand (VoD). This article introduced the survivable *optical internet* from several aspects, which includes a number of selected proposals and implementations for performing protection and restoration mechanisms in order to deal with single failure in network components.

This article first described the evolution of data networks and the importance of *survivability*, then some background knowledge and terminology were presented. Sections 3 and 4 described the latest progress in the protection and restoration strategies proposed by both industry and academia, which includes a comparison between all the approaches mentioned in terms of *scalability*, *dynamicity*, *class of service*, *capacity efficiency*, and *restoration speed*. Table 1 lists the results of the discussion, and compares the strategies mentioned in this article.

BIOGRAPHIES

Hussein Mouftah joined the Department of Electrical and Computer Engineering at Queen's University,

Table 1. Protection versus Restoration Strategies									
	Scalability	Dynamicity	Class of Service	Capacity Efficiency	Restoration Speed				
Ring/node cover	Low	Low	Low	Low	Low				
p-cycle	Low	Low	Low	High	High^b				
Protection cycles	Low	Low	N/A^a	Low	High^a				
Successive Survivable	Medium	High	Low	Medium	Medium				
Routing (SSR)									
Asymmetrically Weighted Survivable	Medium	High	Low	Medium	Medium				
Routing (AWSR)									
Static-SLSP	High^{c}	High^{c}	High^{c}	Medium	Medium				

m 11 1

^aProtection Cycles is a pure link-based protection, in which every link has a *dedicated protection* cycle. ^bThe p-cycle has the highest *restoration speed* because of its preconfigured spare resources instead of being only preplanned.

^cS-SLSP divides a working path into several protection domains, and support dynamic allocation of working protection path pairs, which can achieve scalability, dynamicity, and class of service.

Kingston, Canada, in 1979, where he is now a full professor and the department associate head, after three years of industrial experience mainly at Bell Northern Research of Ottawa (now Nortel Networks). He obtained his Ph.D. in EE from Laval University, Quebec, Canada, in 1975 and his B.Sc. in EE and M.Sc. in CS from the University of Alexandria, Egypt, in 1969 and 1972, respectively. He served as editor in chief of the IEEE Communications Magazine (1995-97) and IEEE Communications Society Director of Magazines (1998-99). Dr. Mouftah is the author or coauthor of two books and more than 600 technical papers and eight patents in this area. He is the recipient of the 1989 Engineering Medal for Research and **Development of the Association of Professional Engineers** of Ontario (PEO). He is the joint holder of a honorable mention for the Frederick W. Ellersick Price Paper Award for Best Paper in Communications Magazine in 1993. He is also the joint holder of two Outstanding Paper Awards; for papers presented at the IEEE Workshop on High Performance Switching and Routing (2002) and at the IEEE 14th International Symposium on Multiple-Valued Logic (1984). He is the recipient of the IEEE Canada (Region 7) Outstanding Service Award (1995). Dr. Mouftah is a fellow of the IEEE (1990).

Pin-Han Ho received his B.S. and M.S. degrees from the Department of Electrical Engineering, National Taiwan University, Taiwan, in 1993 and 1995, respectively. He also received an M.Sc. (Eng) degree from the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada, in 2000. He is currently finishing his Ph.D. degree in the same Department of Electrical and Computer Engineering at Queen's University. He is the joint holder of the Outstanding Paper Award for a paper presented at the IEEE Workshop on High Performance Switching and Routing (2002). His area of interest is optical networking and routing and wavelength assignment in survivable WDM-routed networks.

BIBLIOGRAPHY

- 1. A. S. Tanenbaum, Computer Networks, 2nd ed., Prentice-Hall, 1996.
- 2. W. Stallings, Data & Computer Communications, 6th ed., Prentice-Hall, 2000.
- 3. D. Awduche and G.-S. Kuo, MPLS in broadband IP networks, Tutorial T11, IEEE Int. Conf. Communications (ICC2000), New Orleans, June 22, 2000; http://www.icc00.org/technic/ tutor/thursday.htm.
- 4. Y. Rekhter et al., Cisco systems tag switching architecture overview, RFC 2105.
- 5. E. Rosen, A. Viswanathan, and R. Callon, Multiprotocol Label Switching Architecture, RFC 3031, Jan. 2001.
- 6. S. Makam et al., Framework for MPLS-based recovery, Internet Draft, <draft-makam-mpls-recovery-frmwrk-03.txt>, work in progress, July 2001.
- 7. J. H. Hahm et al., Restoration mechanisms and signaling in optical networks, Internet Draft, <draft-many-opticalrestoration-00.txt>, work in progress, Feb. 2001.
- 8. R. Ramaswami and K. N. Sivarajan, Optical Networks-A Practical Perspective, Morgan Kaufmann Publishers, 1998.
- 9. P. Gadiraju and H. T. Mouftah, Channel protection in WDM mesh networks, IEEE Workshop on High Performance Switching and Routing, Dallas, May 2001, pp. 26-30.
- 10. D. Papadimitriou et al., Inference of shared risk link groups, Internet Draft, <draft-many-inference-srlg-00.txt>, work in progress, Feb. 2001.
- 11. N. Ghani et al., Architectural framework for automatic protection provisioning in dynamic optical rings, Internet Draft, <draft-ghani-optical-rings-01.txt>, work in progress, Sept. 2001.
- 12. M. Kodialam and T. V. Lakshman, Dynamic routing of locally restoration bandwidth guaranteed tunnels using aggregated link usage information, Proc. IEEE, Infocom'01, 2001, pp. 376-385.
- 13. D. Zhou and S. Subramaniam, Survivability in optical networks, IEEE Network 16-23 (Nov./Dec. 2000).

- O. J. Asem, Optimal topologies for survivable fiber optic networks using SONET self-healing rings, *Proc. IEEE GLOBECOM*, 1991, Vol. 3, pp. 2032-2038.
- W. D. Grover, Case studies of survivable ring, mesh and mesh-arc hybrid networks, *Proc. IEEE GLOBECOM*, 1992, Vol. 1, pp. 633-638.
- D. Stamatelakis and W. D. Grover, IP layer restoration and network planning based on virtual protection cycles, *IEEE Journal of Selected Areas in Communications* 18(10): 1938-1949 (Oct. 2000).
- C. Qiao, A high speed protocol for bursty traffic in optical networks, SPIE All-Opt. Commun. Syst. 3230: 79-90 (Nov. 1997).
- P.-H. Ho and H. T. Mouftah, A framework for serviceguaranteed path protection of the optical internet, *optical Network Design and Modeling (ONDM)*, Vienna, Austria, Feb. 2001, pp. MO3.2.1–MO3.2.13.
- P.-H. Ho and H. T. Mouftah, SLSP: A new path protection scheme for the optical Internet, *Optical Fiber Communications (OFC) 2001*, Anaheim, CA, March 2001, pp. TuO1.1-TuO1.3.
- D. Stamatelakis and W. D. Grover, Network Restorability Design Using Pre-configured Trees, Cycles, and Mixtures of Pattern Types, TRLabs Technical Report TR-1999-05, Issue 1.0, Oct. 2000.
- W. D. Grover and D. Stamatelakis, Cycle-oriented distributed preconfiguration: Ring-like speed with mesh-like capacity for self-planning network restoration, *Proc. IEEE Int. Conf. Communications*, 1998, Vol. 1, pp. 537–543.
- W. D. Grover, J. B. Slevinsky, and M. H. MacGregor, Optimized design of ring-based survivable networks, *Can. J. Electric. Comput. Eng.* 20(3): 138-149 (Aug. 1995).
- C. Thomassen, On the complexity of finding a minimum cycle cover of a graph, SIAM J. Comput. 26(3): 675-677 (1997).
- G. Ellinas and T. E. Stern, Automatic protection switching for link failures in optical networks with bidirectional links, *Proc. IEEE GLOBECOM*, 1996, Vol. 1, pp. 152–156.
- G. Ellinas, A. G. Hailemariam, and T. E. Stern, Protection cycles in mesh WDM networks, *IEEE J. Select. Areas Commun.* 18(10): 1924–1937 (Oct. 2000).
- Y. Lie, D. Tipper, and P. Siripongwutikorn, Approximating optimal spare capacity allocation by successive survivable routing, *Proc. IEEE Infocom*'01, 2001, Vol. 2, pp. 699-708.
- 27. P.-H. Ho and H. T. Mouftah, Issues on diverse routing for WDM mesh networks with survivability, 10th IEEE Int. Conf. Computer Communications and Networks (ICCCN'01) (in press).
- R. Bhandari, Survivable Networks: Algorithms for Diverse Routing, The Kluwer International Series in Engineering and Computer Science, Kluwer, Boston, 1999.
- 29. J. Tapolcai et al., Algorithms for asymmetrically weighted pair of Disjoint Paths in survivable networks, *Proc. IEEE 3rd Int. Workshop on Design of Reliable Communication Networks* (DRCN 2001), Budapest, Oct. 2001.
- P.-H. Ho and H. T. Mouftah, Spare Capacity Re-allocation with S-SLSP for the Optical Next Generation Internet, Research Report-01-011, Optical Networking Lab., ECE Dept., Queen's University, Kingston, Ontario, Sept. 2001.
- P. Mateti and N. Deo, On algorithms for enumerating all circuits of a graph, SIAM J. Comput. 5(1): 90-99 (March 1976).

SYNCHRONIZATION IN DIGITAL COMMUNICATION SYSTEMS

M. LUISE U. MENGALI M. MORELLI University of Pisa Pisa, Italy

The word synchronization (often abbreviated sync) refers to signal processing functions accomplished in digital communication receivers to achieve correct alignment of the incoming waveform with certain locally generated references. For instance, in a baseband pulse amplitude modulation system the signal samples must be taken with a proper phase so as to minimize the intersymbol interference. To this purpose the receiver generates clock ticks indicating the location of the optimum sampling times. As a second example, in bandpass transmissions a coherent receiver needs carrier synchronization (or recovery), which means that the demodulation sinusoid must be locked in phase and frequency to the incoming carrier. Clock and carrier recovery are instances of signal synchronization, which is carried out within the physical layer of the system. By contrast, network synchronization is usually performed on digital data streams at the higher layers of the ISO/OSI stack. This happens for example with the time alignment of data frames within the format of the digital stream. As network synchronization is usually easier to implement, we will concentrate to a larger extent on signal synchronization, with a short overview of the former at the end of the article.

Signal synchronization is a crucial issue in digital communication receivers. The advent of low-cost and highpower VLSI circuits for digital signal processing (DSP) has dramatically affected the receiver design rules. At the moment of writing this article the vast majority of newly designed transmission equipment is heavily based on DSP components and techniques.¹ For this reason, we concentrate in the following on DSP-based synchronization, with just occasional references to analog methods. After a brief introduction to the general topic of clock and carrier recovery (Section 1), we consider synchronization for narrowband signals in Section 2 and wideband spread-spectrum signals in Section 3. The case of multicarrier transmission is dealt with in Section 4, and is followed by a short review of frame synchronization in Section 5.

1. AN INTRODUCTION TO SIGNAL SYNCHRONIZATION

The ultimate task of a digital communication receiver is to produce an accurate replica of the transmitted data sequence. With Gaussian channels, the received signal component is completely known except for the data

¹ Exceptions are very-high-speed systems for fiberoptic communications [beyond 1 Gbps (gigabits per second)], wherein the signal processing functions are implemented with analog components.
symbols and a group of variables, denoted *synchronization parameters*, which are ultimately related to the signal time displacement with respect to a local reference in the receiver. Reliable data detection requires good estimates of these parameters. Their measurement is referred to as *synchronization* and represents a crucial part of the receiver. The following examples help illustrate the point.

In baseband synchronous transmission the information is conveyed by uniformly spaced pulses representing bit values. The received signal is first passed through a filter (usually matched to the incoming pulses) and then is sampled at the symbol rate. To achieve the best detection performance, the arrival times of the pulses must be accurately located so that the filtered waveform is sampled at "optimum" times. A circuit implementing this function is called *timing* or *clock synchronizer*.

As a second example, consider a bandpass communication system with coherent demodulation. Optimum detection requires that the received signal be converted to baseband making use of a local reference with the same frequency and phase as the incoming carrier. This calls for accurate frequency and phase measurements, since phase errors may severely degrade the detection process. Circuits performing such measurements, as well as appropriate corrections to the local carrier, are called *carrier synchronizers*.

In the following, we will specifically deal with carrier and clock synchronization of bandpass-modulated signals. The function of clock synchronization for (carrierless) baseband transmission will be seen as a special case that can be derived from bandpass techniques with minor modifications only.

From the observations above it is clear that synchronization plays a central role in the design of digital communication equipment since it greatly affects overall system performance. Also, synchronization circuits represent such a large portion of a receiver's hardware and/or software that their implementation has a considerable impact on the overall cost [1,2].

2. SIGNAL SYNCHRONIZATION IN NARROWBAND TRANSMISSION SYSTEMS

By *narrowband* we mean signals whose bandwidth around their carrier frequency is comparable to the information bit rate. Most popular narrowband-modulated signals are phase shift keying (PSK), quadrature amplitude modulation (QAM), and a few nonlinear modulations such as Gaussian minimum shift keying and the broader class of continuous-phase modulations (CPM). We will restrict our attention to PSK and QAM for the sake of simplicity.

2.1. Signal Model and Synchronization Functions

The mathematical model of a modulated signal is

$$s_{\text{mod}}(t) = s_R(t)\cos(2\pi f_0 t) - s_I(t)\sin(2\pi f_0 t)$$
(1)

where f_0 is the carrier frequency. The quantity $s(t) = s_R(t) + js_I(t)$ is the complex envelope of $s_{mod}(t)$ (where

$$j = \sqrt{-1}$$
, and consists of uniformly spaced pulses

$$s(t) = \sum_{i} c_{i}g(t - iT)$$
⁽²⁾

where c_i is the *i*th transmitted symbol, T is the signaling interval, and g(t) is the basic pulseshape. Symbols $\{c_i\}$ are taken from a *constellation* of M points in the complex plane. With M-PSK modulation, c_i may be written as $c_i = e^{i\alpha_i}$, where $\alpha_i \in \{0, 2\pi/M, \ldots, 2\pi(M-1)/M\}$. With M-QAM signaling we have $c_i = a_i + jb_i$, where a_i and b_i belong to the set $\{\pm 1, \pm 3, \ldots, \pm(\sqrt{M}-1)\}$.²

The physical channel corrupts the information-bearing signal with different impairments such as distortion, interference, and noise. When the only impairment is additive noise, the received waveform takes the form

$$r_{\rm mod}(t) = s_{\rm mod}(t-\tau) + w_{\rm mod}(t) \tag{3}$$

where τ is the propagation delay and $w_{\rm mod}(t)$ is background noise. For the additive white Gaussian noise (AWGN) channel, $w_{\rm mod}(t)$ is a Gaussian random process with zero mean and two-sided power spectral density $N_0/2$. The signal-to-noise ratio (SNR) is defined as the ratio of the power of the signal component to the noise power in a bandwidth equal to the inverse of the signaling interval T(the so-called Nyquist bandwidth).

As is shown in Fig. 1, the demodulation is performed by multiplying (downconverting) $r_{mod}(t)$ by two local quadrature references: $2 \cos(2\pi f_{LO}t + \varphi)$ and $-2 \sin(2\pi f_{LO}t + \varphi)$. The products are lowpass (LP)-filtered to eliminate the frequency components around $f_{LO} + f_0$. In general, the local carrier frequency f_{LO} is not exactly equal to f_0 , and the difference $\nu = f_{LO} - f_0$ is referred to as *carrier frequency offset*. Assuming that the filters' bandwidth is large enough to pass the signal components undistorted, and collecting the filter outputs into a single complex-valued waveform $r(t) = r_R(t) + jr_I(t)$, after some mathematics it is found that

$$r(t) = e^{j(2\pi\nu t+\theta)} \sum_{i} c_i g(t-iT-\tau) + w(t)$$
(4)



Figure 1. I/Q signal demodulation.

² Here, the number of points in the constellation is an *even* power of 2. Different constellations with M equal to an *odd* power of 2 or with an arbitrary number of points are commonly used as well.

where $\theta = -(2\pi f_0 \tau + \varphi)$ and $w(t) = w_R(t) + jw_I(t)$ is the noise contribution. Inspection of this equation reveals that the demodulated signal contains the following unknown parameters: the *frequency offset* v, the *phase offset* θ , and the *timing offset* τ . As is now explained, reliable data detection cannot be achieved without knowledge of these parameters.

Assume that the frequency offset is much smaller than the signal bandwidth, as is often the case. Then, passing r(t) through a filter matched to g(t) produces

$$x(t) = e^{j(2\pi vt + \theta)} \sum_{i} c_i h(t - iT - \tau) + n(t)$$
(5)

where n(t) is filtered noise and h(t) is the convolution of g(t) with g(-t). To single out the effects of frequency and phase errors, we assume for the moment that τ is perfectly known and that h(t) satisfies the first Nyquist criterion for the absence of intersymbol interference (ISI):

$$h(kT) = \begin{cases} 1 & \text{for } k = 0\\ 0 & \text{for } k \neq 0 \end{cases}$$
(6)

Then, sampling x(t) at the ideal clock instants $t_k = kT + \tau$, we get

$$x[k] = c_k e^{j[2\pi \nu (kT + \tau) + \theta]} + n[k]$$
(7)

where n[k] is the noise sample. From (7) it is seen that the useful component of x[k] is rotated by a time-varying angle $\psi[k] = 2\pi v(kT + \tau) + \theta$ with respect to its correct position, and this may have a catastrophic effect on system performance. For instance, consider a simple binary PSK signal with $c_k \in \{-1, +1\}$. Regeneration of the digital datastream is carried out making the decision $\hat{c}_k = \pm 1$ according to whether $\Re e\{x[k]\} = \pm 1$. Neglecting the noise for simplicity, it is easily found that $\Re e\{x[k]\} = c_k \cos(\psi[k])$, which means that the receiver makes decision errors whenever $\pi/2 < |\psi[k]| \le \pi$. This condition is certainly met (sooner or later) in the presence of a frequency offset, but it may also hold true due to a *phase* offset even when the frequency offset is negligible.

Now consider a timing error. In doing so we assume that frequency and phase offsets have already been compensated for, and that the receiver has elaborated an *estimate* $\hat{\tau}$ of the channel delay (in general, $\hat{\tau} \neq \tau$). Then, sampling the matched-filter output at the clock instants $\hat{t}_k = kT + \hat{\tau}$ yields

$$x[k] = c_k h(-\varepsilon) + \sum_{m \neq 0} c_{k-m} h(mT - \varepsilon) + n[k]$$
(8)

where $\varepsilon = \tau - \hat{\tau}$ is the *timing error*. Note that, as ε is usually small compared to the sampling period *T*, we have $h(-\varepsilon) \approx 1$. The first term on the right-hand side (RHS) of (8) is the useful component, the second is a disturbance referred to as *intersymbol interference* (ISI), and the third is Gaussian noise. Clearly, the ISI tends to "mask" the useful component and can produce decision errors even in the absence of noise. For this reason h(t) is usually given a shape satisfying the first Nyquist criterion. This would make the ISI vanish if x(t) were sampled exactly at the nominal clock instants $t_k = kT + \tau$ (i.e., $\varepsilon = 0$). In practice



Figure 2. Coherent receiver with sync functions.

some residual ISI is inevitable, but its amount can be limited by keeping ε small.

To sum up, Fig. 2 illustrates the signal synchronization functions of a coherent receiver for PSK or QAM signals. The blocks indicated as frequency, phase, and timing recovery have the task of providing reliable estimates $\hat{\nu}, \hat{\theta}, \hat{\eta}$ and $\hat{\tau}$ of the corresponding sync parameters ν, θ , and τ , respectively. With such estimates, the receiver compensates for the sync offsets. To do so, it first multiplies (downconverts) the received signal by $e^{-j2\pi \hat{v}t}$; next it takes samples at the correct clock instants $\hat{t}_k = kT + \hat{\tau}$, and finally it multiplies the signal samples by $e^{-j\hat{ heta}}$ to cancel out the residual phase offset. It is worth noting that Fig. 2 has only illustration purposes; indeed, the real receiver architecture may be somewhat different. For instance, phase recovery may be accomplished before matched filtering and without exploiting timing information. Similarly, timing recovery may be derived directly from the demodulated waveform r(t), prior to frequency compensation.

So far our discussion has concentrated on the simple case of transmission on the AWGN channel, which is typical of satellite communications. Unfortunately this simplified scenario does not fit many modern communication systems, as, for example, mobile cellular networks or high-speed wireline systems for the access network. In such systems the channel frequency response is not flat across the signal bandwidth and the signal undergoes unpredictable linear distortions. This implies that, in addition to the sync parameters mentioned above, the receiver also has to estimate the *channel impulse response* (CIR). Indeed, the received signal takes the form

$$r(t) = e^{j2\pi v t} \sum_{i} c_i p(t - iT) + w(t)$$
(9)

where p(t), the (complex-valued) CIR to be estimated, may be modeled as the convolution of three impulse responses: those of the transmit filter, the physical channel, and the receive filter. Comparing (9) to (4), it is seen that the parameters θ and τ are no longer visible as they have been incorporated into p(t). Thus, CIR estimation has become an augmented synchronization problem. Although CIR estimation and synchronization functions tend to overlap and possibly merge in modern communication equipment, in the following we stick to synchronization issues only. The interested reader is referred to articles on estimation and equalization chapters elsewhere in this encyclopedia.

A further warning about Fig. 2 is that the diagram seems to suggest that synchronization is always derived from the information-bearing signal (self-synchronization). Actually, the transmitted signal often contains known sequences periodically inserted into the datastream. They are referred to as *preambles* or *training sequences* and serve to ease the estimation of the sync parameters. Notwithstanding, in other cases synchronization is carried out by exploiting a dedicated channel, the socalled pilot channel (pilot-aided synchronization). Even though the method requires extra bandwidth, it is often adopted whenever channel distortion and multiple access interference make the detection process critical. This occurs with third-generation cellular systems (UMTS in Europe and Japan, and cdma2000 in the Americas) where both the downlink (from the radio base station to the mobile phones) and the uplink do employ pilot channels.

2.2. Performance of Synchronization Functions

From the discussion in the previous section it is recognized that synchronization functions may be split into two parts: *estimation* of the sync parameters and *compensation* of the corresponding estimated offsets. Strategies to accomplish these tasks are now pinpointed, and the major performance indicators are introduced.

The estimation of a sync parameter is accomplished by operating on the digitized samples of the received signal. Methods to do so fall into two categories: either feedforward (open loop) or feedback (closed loop). The former provide estimates based on the observation of the received signal over a finite time interval (observation window), say, $0 \le t \le T_0$. At the end of the interval the estimate is released and is used for compensation of the relevant offset. Compensation is applied either back on the stored signal samples within the observation window (if adequate digital memory is available), or on the samples of the subsequent window (assuming that the sync parameters are almost constant from one window to the next). An attractive feature of feedforward schemes is that they have short estimation times and, as such, are particularly suited for burst-mode transmissions where fast synchronization is mandatory. Feedback schemes, on the other hand, have much longer estimation times as they operate in a recursive fashion. For instance, in a clock recovery circuit the timing estimate is periodically updated (usually at symbol rate) according to an equation of the type

$$\hat{\tau}[k+1] = \hat{\tau}[k] - \gamma e_{\tau}[k] \tag{10}$$

where $\hat{\tau}[k]$ is the estimate at the *k*th step, $e_{\tau}[k]$ is an *error signal*, and γ is a design parameter (*step size*). The error signal $e_{\tau}[k]$ is provided by the synchronizer and (to a first approximation) is proportional to the *k*th estimation error $\hat{\tau}[k] - \tau$. Equation (10) is reminiscent of a feedback control system; since the correction term $-\gamma e_{\tau}[k]$ has a sign opposite the error $\hat{\tau}[k] - \tau$, the latter is steadily forced toward zero. Feedback synchronizers are akin to the analog phase-locked loops (PLLs) used for carrier acquisition and tracking. Feedforward synchronizers have no counterparts in analog hardware.

The estimate of a sync parameter λ is a random variable depending on samples of the received signal. Accordingly, it is customary to qualify the accuracy of an estimate by specifying its *mean value* and *mean-square error* (MSE). The estimate $\hat{\lambda}$ is said to be *unbiased* if its mean value equals the true value λ . The bias is defined as

$$b(\hat{\lambda}) = E\{\hat{\lambda}\} - \lambda \tag{11}$$

2475

where $E\{\cdot\}$ means statistical expectation. In general we would like $b(\hat{\lambda})$ to be zero since this means that the estimates would coincide with λ at least "on average." The MSE of the estimator is given by

$$MSE(\hat{\lambda}) = E\{(\hat{\lambda} - \lambda)^2\}$$
(12)

Clearly, the smaller $MSE(\hat{\lambda})$ is, the more accurate the estimate $\hat{\lambda}$ is, in that the latter will be "close" to λ with high probability.

In searching for good estimators, one wonders what is the ultimate accuracy that can be achieved. An answer is given by the Cramér-Rao bound (CRB) [3], which represents a lower limit to the MSE of *any unbiased estimator*:

$$MSE(\hat{\lambda}) \ge CRB(\hat{\lambda})$$
(13)

Knowledge of the CRB is very useful as it establishes a benchmark against which the performance of practical estimators can be compared.

Another fundamental performance parameter of a sync system is the *acquisition time*. With feedforward schemes this is just the time needed to compute the estimate $\hat{\lambda}$ and coincides with the length of the observation window (plus possibly some extra time to carry out the calculations). With feedback systems, however, the acquisition time represents the number of iterations needed to achieve a steady-state estimation condition. As is intuitively clear from Ref. 10, the acquisition time depends on the step size γ . In general, the larger γ is, the quicker the convergence. However, increasing γ degrades the estimation accuracy in the steady state. In practice, some tradeoff between these conflicting requirements must be sought.

A possible drawback of closed-loop schemes is the *cycle* slipping phenomenon [4]. Briefly, imagine a steady-state condition in which $\hat{\lambda}[k]$ is fluctuating around the true value λ . Fluctuations are usually small but, occasionally, they grow large as a consequence of the combined effects of noise and the random nature of the data symbols. If a large deviation occurs, it may happen that $\hat{\lambda}[k]$ is "attracted" toward an equilibrium point other than the initial steady-state value [1]. When this happens, a synchronization failure (*cycle slip*) occurs, and we say that the loop has lost lock. Cycle slips must be rare events in a well-designed loop because they reflect the presence of large errors. Their rate of occurrence is usually measured by the *mean time to lose lock*, specifically, the average time between two consecutive sync loss events.

2.3. Frequency Synchronization

Feedforward frequency synchronizers are typically used in burst-mode transmissions (as in satellite time-division multiple access), with frequency offsets much smaller than the signaling rate. In these circumstances timing recovery is performed first and the frequency estimator operates on symbol rate samples. Most frame formats are endowed with a *preamble* that is used to cancel the modulation in the samples (7) from the matched filter by dividing x[k] by c_k . This *Data-Aided* (DA) operation produces

$$z[k] = e^{j[2\pi\nu(kT+\tau)+\theta]} + n'[k] \quad 0 \le k \le N-1$$
(14)

where $n'[k] = n[k]/c_k$ and N is the observation length in symbol intervals. When no preamble is available, the data modulation is usually wiped out by processing x[k] through a nonlinear function. With *M*-ary PSK, for example, x[k] is raised to the *M*th power. This is called *non-data-aided* (NDA) or "blind" processing.

Several algorithms have been devised to estimate ν from z[k]. An efficient method, proposed by Rife and Boorstyn (RB) [5], consists of computing the Fourier transform (FT) of the sequence z[k]

$$Z(f) = \sum_{k=0}^{N-1} z(k) e^{-j2\pi f kT}$$
(15)

and taking $\hat{\nu}$ as the value of f where |Z(f)| achieves a maximum. It turns out that the RB algorithm is unbiased and attains the Cramér-Rao bound

$$CRB_{\nu} = \frac{3}{2\pi^2 T^2 N (N^2 - 1)} \times (SNR)^{-1} (Hz)^2$$
 (16)

at intermediate to high signal-to-noise ratios. At low SNRs, however, the estimates are plagued by occasional large estimation errors (*outliers*) and the variance of the estimator exhibits a *threshold effect*, typical of nonlinear estimation schemes. The threshold effect manifests as an abrupt increase of the estimation MSE as the SNR decreases below a certain value (the estimator's threshold).

Although the RB algorithm can be implemented through efficient fast Fourier transform (FFT) techniques, its complexity is too high with large data records. Simpler methods based on linear regression of $\arg\{z[k]\}$ have been proposed in [6,7]. Their main drawback is that their threshold may be as high as 7–8 dB, and cannot be lowered by increasing the observation length (as occurs with the RB scheme). Other estimators are discussed in [8,9] and are based on the sample correlation function of z[k]

$$R[m] = \frac{1}{N-m} \sum_{k=m}^{N-1} z[k] z^*[k-m] \quad 1 \le m \le Q$$
(17)

where Q is a design parameter. In general, increasing Q improves the accuracy of the estimates but reduces the maximum offset that can be estimated (*estimation range*). A method to make the estimation range independent of Q is given elsewhere [10].

The frequency estimators described in other studies [5-10] were conceived specifically for unmodulated signals (or for DA operation), but they can also be used in NDA recovery [11]. The price to pay is an increase of the threshold with respect to DA operation, especially with large-signal constellations. This may be a serious drawback with efficient forward error-correcting codes, like Turbo codes, where the operating SNR is as low as 1-2 dB.

All of the above frequency synchronizers produce biased estimates when the transmitted signal undergoes linear distortion, as happens in wideband transmissions. The issue of frequency recovery for selective channels is difficult to deal with. Suffice it to say that, even theoretically, it is not clear whether unbiased frequency estimates can be obtained without knowledge of the CIR. The problem of joint estimating CIR and frequency offset is addressed in Ref. 12 assuming that a preamble is available. The resulting scheme performs well but is complex to implement. A different solution [13] exploits knowledge of the channel statistics.

A mixed analog/digital feedback frequency synchronizer is sketched in Fig. 3. As is seen, the frequencycorrected signal $r'(t) = r(t)e^{-j\phi(t)}$ is fed to a *frequency error detector* (FED) whose purpose is to generate the error signal $e_v[k]$. The latter gives an indication of the difference between the current estimate $\hat{v}(k)$ from the digitally controlled oscillator (DCO) and the true value v. Similar to (10), the error signal is then used in a feedback loop to update the estimates in a recursive fashion:

$$\hat{\nu}[k+1] = \hat{\nu}[k] - \gamma e_{\nu}[k] \tag{18}$$

The frequency estimate is then used in the DCO to generate the exponential $e^{-j\phi(t)}$ such that

$$\frac{d\phi(t)}{dt} = 2\pi\,\hat{\nu}[k], \quad kT \le t < (k+1)T \tag{19}$$

Compared with feedforward schemes, feedback synchronizers typically have larger acquisition times but can track large slow time-varying frequency offsets. For this reason they are well suited for continuous-mode transmissions. Their counterpart in analog hardware is the traditional automatic frequency control (AFC) loop, commonly used in radio receivers.

The heart of closed-loop schemes is the FED. The maximum-likelihood-based FED [14], the quadricorrelator [15] and the dual-filter detector [16], all share the same kind of loop error signal:

$$e_{\nu}[k] = \int_{kT}^{(k+1)T} \Im m\{y^*(t)z(t)\}dt$$
 (20)

where y(t) and z(t) are obtained by passing r'(t) through two suitable lowpass filters, $\Im m\{\cdot\}$ means *imaginary part*



Figure 3. Feedback frequency recovery.

of the enclosed quantity, and the asterisk denotes complex conjugation. In its simplest version, y(t) is just the matched filter output and z(t) is a replica of y(t) delayed by a symbol interval, namely, z(t) = y(t - T). The preceding schemes do exploit neither knowledge of data symbols nor timing information and can recover frequency offsets as large as the symbol rate 1/T. Unfortunately, their accuracy is far from the CRB and their estimates are biased in the presence of frequency-selective fading. The reason of the bias is that they take the center of gravity of the received spectrum as an estimate of the carrier frequency. With multipath propagation, however, the signal spectrum is distorted by the channel frequency response, and its center of gravity is moved off its original position. A closed-loop frequency synchronizer for transmissions over multipath channels has been proposed [17]. It gives unbiased estimates provided that the channel is flat over the rolloff regions of the received spectrum.

2.4. Phase Synchronization

When using phase-coherent detection the phase offset must be properly compensated for before data detection takes place. In a DSP-based receiver, phase recovery is usually performed after timing correction using the symbol rate samples x[k] from the matched filter. Information symbols may be known (training sequence) or not. Assuming that the frequency offset is zero or has been perfectly compensated for [i.e., setting v = 0 in (7)], we have

$$x[k] = e^{j\theta}c_k + n[k] \tag{21}$$

When the symbols c_k are known, a feedforward phase synchronizer yields the maximum-likelihood (ML) estimate of θ as follows:

$$\hat{\theta} = \arg\left\{\sum_{k=0}^{N-1} c_k^* x[k]\right\} - \pi \le \theta < \pi$$
(22)

where N is the observation length in symbol intervals and $\arg\{\cdot\}$ denotes the argument of the enclosed complex value. The estimate (22) is unbiased and achieves the CRB

$$CRB_{\theta} = \frac{1}{2N} (SNR)^{-1} (rad)^2$$
(23)

When no preamble is available, data modulation can be removed from x[k] by means of suitable nonlinear processing. For example, with *M*-ary PSK, x[k] is raised to the *M*th power to yield the following open-loop NDA phase estimate:

$$\hat{\theta} = \frac{1}{M} \arg\left\{\sum_{k=0}^{N-1} x^M[k]\right\} - \frac{\pi}{M} \le \theta < \frac{\pi}{M}$$
(24)

A variant of this estimator is proposed by Viterbi and Viterbi (VV) in [18]. The VV algorithm converts x[k] to polar coordinates $x[k] = \rho[k]e^{i\varphi[k]}$ and then replaces $x^{M}[k]$ in (24) with either $\rho^{2}[k]e^{iM\varphi[k]}$ or $e^{iM\varphi[k]}$ (depending on the SNR and the number M of constellation points). Although this nonlinear processing degrades the

estimation accuracy, the VV algorithm achieves the CRB at high SNR (i.e., asymptotically).

2477

As indicated in (24) the NDA schemes are forced to operate on a smaller phase interval than the whole 2π angle, namely, $-\pi/M \le \theta < \pi/M$. This implies that they give estimates that are ambiguous by multiples of $2\pi/M$. Distinct phase offsets differing by $2\pi/M$ are "mapped" onto the same estimated value within the base interval $-\pi/M \le \theta < \pi/M$. The ambiguity is usually resolved by means of a *unique word* [1] or by differential encoding and decoding [19].

NDA phase estimators for QAM constellations are obtained by letting M = 4 in (24) (fourth power estimators) or in a VV-like algorithm. The estimation accuracy of such schemes tends to fall short of the CRB (even at high SNRs) as the number of constellation points increases [20].

The phase recovery algorithms illustrated so far have all a feedforward structure that makes them suited to burst-mode transmission. However, feedback loops are preferred with continuous transmissions for the sake of simplicity. Probably the most popular feedback scheme is illustrated in Fig. 4. Here the phase-compensated signal samples $y[k] = x[k]e^{-j\hat{\theta}[k]}$ and the detected symbols \hat{c}_k are fed to the phase error detector (PED) to generate the error signal

$$e_{\theta}[k] = \Im m\{y^*[k] \cdot \hat{c}_k\}$$
(25)

The latter is then passed through an IIR filter which updates the phase estimate according to

$$\hat{\theta}[k+1] = \hat{\theta}[k] - \gamma e_{\theta}[k]$$
(26)

where γ is the step size. Finally, the lookup table produces the map $\hat{\theta}[k] \rightarrow e^{-j\hat{\theta}[k]}$.

As the PED uses data decisions to build up the error signal, the loop is said to operate in a *decision-directed* (DD) mode. When a preamble is available it may also operate in a data-aided (DA) "training mode" to ease initial acquisition. In this case data decisions \hat{c}_k in (25) are replaced by the preamble symbols.

In the presence of an uncompensated residual frequency error the phase estimates $\hat{\theta}[k]$ are biased if the loop filter is first-order as indicated in (26). To cope with moderate-frequency errors, second-order loop filters are adopted [1]. The phase recovery algorithm (25)–(26) is often referred to as "digital Costas loop" since it resembles



Figure 4. Digital Costas loop for phase recovery.

the analog PLL invented in the 1950s by Costas to track the chrominance subcarrier of color television signals [21].

The steady-state performance of Costas loops is good but, as happens with the NDA feedforward estimators, the estimates are ambiguous by multiples of $2\pi/M$ with *M*-ary PSK and by multiples of $\pi/4$ with QAM modulation. Again, the ambiguity can be resolved either with unique words or with differential encoding and decoding.

In the previous discussion we have concentrated on uncoded transmissions. With trellis-coded modulations the tracking performance of a conventional Costas loop may fail as a result of the decision delay inherent in the detection process. In these circumstances persurvivor-processing (PSP) techniques [22] are preferable. In practice, each surviving path in the trellis diagram generates a phase estimate based on its own tentative decisions. The estimate is then used to extend that survivor one step further in the trellis. The issue of phase recovery with large decoding delays (like those experienced in Turbo decoding [23]) is still an open problem.

2.5. Timing Synchronization

As with the other sync functions, timing recovery consists of two distinct operations: (1) estimation of the timing offset τ (*timing estimation*) and (2) application of the estimate to the sampling process (*timing compensation*).

Two different timing recovery architectures are possible. Figure 5 shows a feedforward scheme with asynchronous signal sampling. The received signal is passed through an antialias filter (AAF) and it is then fed to the A/D converter (represented by the sampler in Fig. 5). The latter is controlled by a free-running oscillator, having no reference whatsoever with the clock of the data signal. The sampling rate $1/T_s$ is usually higher than the symbol rate by an oversampling factor >2. Timing correction is achieved by *interpolating* the samples $x[m] = x(mT_s)$ from the matched filter according to the estimates of τ . This "resynthesizes" signal samples at the correct timing instants, which are seldom present in the digitized stream x[m]. Simple piecewise polynomial interpolators are described by Erup et al. [24].

The alternative architecture (see Fig. 6) involves *synchronous signal sampling* and is particularly useful with high-data-rate modems where oversampling is too expensive or not feasible. Here, timing correction is accomplished through a feedback loop wherein a timing error detector (TED) drives a numerically controlled oscillator (NCO). The latter updates the timing estimates



Figure 5. Feedforward timing recovery with asynchronous sampling.



Figure 6. Feedback timing recovery with synchronous sampling.

according to

>

$$\hat{\tau}[k+1] = \hat{\tau}[k] - \gamma e_{\tau}[k] \tag{27}$$

where $e_{\tau}[k]$ is the TED output and γ is the step size. In so doing, the NCO produces a sequence of clock pulses that are *synchronous* with the data symbol clock.

Feedback timing estimation may be performed jointly with carrier phase recovery in a DD mode. To this end the zero-crossing detector [15] or the Muller-Mueller detector [25] may be employed. The main drawback of these methods is that spurious locks may occur with large QAM constellations due to complex interactions between phase and timing loops. A simpler approach is to recover timing independently of the carrier phase. The Gardner detector (GD) [26] and the NDA early-late detector (ELD) [27] are widely used for timing recovery in the absence of phase information. They operate with an oversampling factor of 2 and generate the following error signals

$$e_{\rm GD}[k] = \Re e\{(x(kT + \hat{\tau}[k]) - x(kT - T + \hat{\tau}[k-1])) \\ \times x^*(kT - T/2 + \hat{\tau}[k-1])\}$$
(28)

$$e_{\rm ELD}[k] = \Re e\{(x(kT - T/2 + \hat{\tau}[k-1]) - x(kT + T/2 + \hat{\tau}[k]))$$

$$\langle x^*(kT + \hat{\tau}[k])\} \tag{29}$$

where x(t) is the output of the matched filter. Note that both $e_{\text{GD}}[k]$ and $e_{\text{ELD}}[k]$ are insensitive to phase errors, since any possible phase offset term $e^{j\theta}$ on the received signal vanishes in the product between the samples of signal x(t) and its own complex conjugate.

Returning to feedforward schemes, the most popular algorithm in this class is that proposed by Oerder and Meyr (OM) [28]. It needs an oversampling factor M greater than 2, and has the form

$$\hat{\tau} = -\frac{T}{2\pi} \arg \left\{ \sum_{k=0}^{MN-1} |x(kT_s)|^2 e^{-j2\pi k/M} \right\}$$
 (30)

where N is the observation length (in symbol intervals). Simulations indicate that both GD and the OM algorithms have good performance with raised-cosine-shaped pulses, provided that the rolloff factor is sufficiently large. The estimation accuracy becomes poor as the signal bandwidth decreases.

The above algorithms are tailored for transmissions over the AWGN channel. With multipath channels the situation is more complex. For example, the MSE at the output of a symbol-spaced equalizer is very sensitive



Figure 7. Coherent CDMA receiver.

to the timing phase and, for a given CIR, an optimal phase exists that minimizes the bit error rate (BER). The key point is how this phase can be found and properly tracked as the channel varies in time. A reasonable solution is proposed by Godard [29] under some restrictive assumptions. Timing recovery may be avoided by resorting to fractionally spaced equalizers whose performance is insensitive to the sampling phase [30].

2.6. Timing Synchronization in Baseband Transmission

The timing algorithms for bandpass signals described in Section 2.5 can be used with minor changes for baseband transmission as well. The GD and ELD are directly translated to baseband by just interpreting x(t) as a real-valued signal and consequently by dropping the complex-conjugate operation. The same is true with the OM open-loop estimator. It is worth noting that the ELD in (29) is just the digital counterpart of the popular analog early-late synchronizer for rectangular data pulses [31].

3. SYNCHRONIZATION FOR SPREAD-SPECTRUM AND CDMA SIGNALS

3.1. Signal Model and Synchronization Functions

With the term *spread-spectrum* we address any modulated signal whose bandwidth around the carrier frequency is much larger than its information bit rate. The most popular spread-spectrum format for commercial applications is *direct-sequence* spread-spectrum (DS/SS) wherein spectral spreading is directly obtained through the use of a high-rate sequence (the code) of binary values called chips. DS/SS is the basis for code-division multiple access (CDMA) techniques adopted in mobile cellular systems such as IS95, cdma2000, and UMTS. In a CDMA network, each user is assigned a specific code (or signature sequence). It consists of a pseudorandom sequence with a repetition period of L chips, and with a chip rate $1/T_c$ equal to N times the symbol rate 1/T. The parameter N is called the *spreading factor* since the signal bandwidth is about $1/T_c$ and is N times larger than the bandwidth of the modulating signal, 1/T.

Consider the *i*th user in a CDMA network and call $b_k^{(i)}$ his/her binary datastream, with index k running at the symbol rate 1/T. Denoting by $\lfloor n/N \rfloor$ the integer part of n/N and by $|n|_N$ the remainder of the division, the baseband equivalent signal transmitted by the *i*th user takes the form

$$s^{(i)}(t) = \sum_{\ell} c^{(i)}_{|\ell|_L} b^{(i)}_{\lfloor \ell/N \rfloor} g(t - lT_c)$$
(31)

where g(t) is the chip pulseshape and $\{c_{\ell}^{(i)}; 0 \le \ell \le L-1\}$ is the signature sequence (with ℓ ticking at the chip rate).

Figure 7 is a block diagram of a conventional coherent receiver for user 1. Waveform r(t) is the sum of signals from all the active users (in number of *I*). Assuming an AWGN channel, it takes the form

$$r(t) = e^{j(2\pi\nu_1 t + \theta_1)} s^{(1)}(t - \tau_1) + \sum_{i=2}^{I} A_i e^{j(2\pi\nu_i t + \theta_i)} s^{(i)}(t - \tau_i) + w(t)$$
(32)

where w(t) is the noise; v_i and θ_i are carrier frequency and phase offsets, respectively; while A_i and τ_i account for the attenuation and delay experienced by the *i*th signal. We say that the CDMA system is *synchronous* when all signals share the same chip and symbol framework: $\tau_i = \tau$ for i = 1, 2, ..., I. This happens for example in the downlink of a mobile communication network. Otherwise, when the signals are not bound by synchronicity constraints (as in the case of the uplink from mobiles to base station), we speak of *asynchronous* multiple access.

Returning to Fig. 7, after frequency and phase correction, the received waveform is fed to the chip matched filter (CMF) and then is sampled at chip rate. Spectral despreading is performed by multiplying the samples $x[\ell]$ by a locally generated replica of the user's code and, finally, the resulting sequence is accumulated over a symbol period.

Correct receiver operation requires accurate recovery of the signal time offset τ_1 to ensure that the local code replica is properly aligned with the signature sequence in the received signal. This goal is usually achieved in two steps—a coarse alignment is obtained first and is then used as a starting point for code tracking.

Assuming perfect carrier recovery and code alignment, it turns out that the decision variable $z^{(1)}[k]$ at the accumulator output is given by

$$z^{(1)}[k] = b_k^{(1)} + \eta_{\text{MAI}}^{(1)}[k] + n[k]$$
(33)

where n[k] is the noise contribution while $\eta_{\text{MAI}}^{(1)}[k]$, the *multiple-access interference* (MAI), accounts for the presence of the other users. MAI can be reduced to zero in synchronous CDMA by using *orthogonal* code sequences (Walsh–Hadamard). With asynchronous CDMA, however, orthogonality cannot be enforced even with Walsh–Hadamard sequences because of the different time offsets τ_1, \ldots, τ_l . In most cases MAI is the major limiting factor to achieve accurate synchronization and reliable data detection in CDMA systems. Approximating MAI as Gaussian noise, the last two terms on the RHS of (33) can be lumped together to give

$$z^{(1)}[k] = b_k^{(1)} + n'[k]$$
(34)

where n'[k] is zero-mean Gaussian with a variance equal to the sum of the variance of n[k] and of the MAI term $\eta_{\text{MAI}}^{(1)}[k]$.

From the discussion above it appears that the synchronization problem in CDMA systems is similar to that encountered with narrowband signals, with two main differences: (1) the presence of MAI and (2) the broader signal bandwidth that makes the time offset compensation (code synchronization) much more complex, as discussed in Section 3.2.

3.2. Code Synchronization

As mentioned earlier, signal demultiplexing relies on the availability at the receiver of a time-aligned version of the spreading code. The delay τ_1 must be estimated and tracked to ensure such an alignment. The main difference with respect to narrowband modulations is the estimation accuracy. In narrowband systems, timing errors must be small compared to the *symbol time* whereas in spread-spectrum systems they must be small compared with the *chip* time, which is N times smaller.

Assume again an AWGN channel and, for simplicity, that the spreading factor equals the code repetition length (the so-called *short-code* DS/SS format). Then, code acquisition amounts to finding the start of the symbol interval and is usually performed by correlating the input signal with the local replica of the code sequence (*sliding correlator*). The magnitude of the correlation is used in two ways: it is compared to a threshold to decide whether the intended user is actually transmitting and, if this is the case, to locate the position of the maximum. This location is taken as a coarse estimate of the delay τ_1 .

The search for the maximum of the correlation may be performed with either serial or parallel schemes [32]. Serial schemes test all the possible code delays in sequence until the threshold is crossed. They are simple to implement but are inherently time-consuming and their acquisition time cannot be established a priori (it can only be predicted statistically). Parallel schemes look for all the possible code epochs in parallel and choose the one corresponding to the maximum correlation. They guarantee short acquisitions but are computationally intensive.

A simplified scheme of a serial detector is sketched in Fig. 8. The PN-code generator provides a spreading code with a tentative initial epoch δ , which is correlated with the (digitized) received signal on a symbol period. The result is squared to cancel out any possible phase offset (noncoherent processing); then it is smoothed on a W-symbol dwell time, and finally it is compared to a threshold. If the threshold is crossed, the code epoch is frozen and the fine timing recovery process is started. Otherwise, the PN-code generator is advanced by one chip and a new trial acquisition is performed. The value of the threshold λ is a design parameter. Low values of λ correspond to high probabilities of false acquisition events caused by large noise peaks. On the other hand, high threshold values produce occasional acquisition failures. The false-alarm probability and missed-detection probability depend on λ and on the SNR. The threshold value is chosen on the basis of the operating conditions.

Conventional correlation-based methods have satisfactory performance in a power-controlled system, when signals from different users arrive at the receiver with comparable amplitudes. However, they fail in a *near-far* situation where strong-powered users interfere with weaker ones. In these cases MAI becomes a serious impairment to achieve accurate code synchronization. Improvements are obtained by taking into account the statistical properties of the MAI. For example, near-far resistant code acquisition is achieved by modeling MAI as colored Gaussian noise [33].

Initial code acquisition provides the receiver with a coarse estimate of the delay τ_1 . Fine timing recovery (also called *code tracking*) is then needed to locate the optimum chip-rate sampling instants. Code tracking is typically performed by means of feedback loops, where a suitable timing error signal is used to update the timing estimate at symbol rate according to an equation of the type (27). Figure 9 depicts the architecture of a digital



Figure 9. Digital delay-lock loop for fine timing recovery of a DS/SS or CDMA signal.



Figure 8. Serial code acquisition of a DS/SS signal.

delay-lock loop (DLL) [34] performing such a function. It is seen that data demodulation relies on the "on-time samples," those taken at the optimum sampling instants. The DLL also needs the "early/late samples," those taken midway between two consecutive on-time samples. The estimate $\hat{\varepsilon}[k]$ of the normalized chip delay ε is then used to drive an NCO or an interpolator, depending on whether synchronous or asynchronous sampling is employed. Moeneclaey and De Jonghe [35] give some examples of low-complexity chip-timing error detectors (CEDs). For instance, detectors insensitive to the phase offset (not requiring prior phase recovery) are expressed by

$$e_1[k] = \Re e\{(z_E[k] - z_L[k])z^*[k]\}$$
(35)

$$e_2[k] = |z_E[k]|^2 - |z_L[k]|^2$$
(36)

The first is reminiscent of the early-late timing detector for narrowband transmissions; the second is typical of spread-spectrum signals. Their performance is satisfactory in the absence of multipath and near-far effects.

The foregoing discussion has been concerned with DS-CDMA transmissions over the AWGN channel. On the other hand, third generation CDMA-based systems are expected to support multimedia services with data rates up to 2 Mbps. In these circumstances the transmission channel is characterized by multipath propagation. A nice feature of CDMA systems is that they can resolve multipath components and optimally combine them by means of a RAKE receiver [19] or other more sophisticated receiver structures based on multiuser detection [36]. In all cases, accurate estimates of the relative delays, amplitudes, and phases of the propagation paths are needed to achieve reliable data detection [37-39]. In particular, the problem of measuring the propagation delays seems crucial since performance degrades rapidly with timing misalignments in excess of a small fraction of the chip interval.

3.3. Carrier Frequency and Phase Synchronization

Code timing recovery is typical of CDMA signals. Carrier recovery, on the other hand, does not differ much with respect to narrowband modulations. In general, carrier recovery is performed after code acquisition, exploiting the despread signal z[k] from the correlator in Fig. 9. A notable exception occurs when the frequency offset is comparable with the symbol rate. In this case code acquisition becomes unreliable because the frequency offset *decorrelates* the signal within the integration window. Indeed, things go as if the signal power were reduced by a factor

$$L = \left[\frac{\pi \nu T_i}{\sin(\pi \nu T_i)}\right]^2 \tag{37}$$

where T_i is the window length. For example, a frequency offset of $1/(2T_i)$ causes a loss of more than 6 dB. On the other hand, frequency offset cannot be reliably estimated unless the code sequence is coarsely acquired. A chicken-egg problem arises that can be approached with *joint* estimation of the frequency offset and the code phase. This leads to a *bidimensional* grid search in which the frequency uncertainty range is partitioned into a number of "bins" and the code acquisition test is repeated for all the bins [32]. In the end, coarse estimates of code phase and frequency offset are available. The latter is then used to correct the local oscillator so as to reduce the residual frequency error to a small fraction of the symbol rate. At that stage frequency tracking is performed by means of conventional feedback schemes such as a quadricorrelator [15] or a dual-filter detector [16].

Frequency recovery for CDMA transmissions on frequency-selective channels is still an open problem. Current research investigates methods to alleviate the combined effects of MAI and multipath on the acquisition process.

4. SYNCHRONIZATION IN MULTICARRIER TRANSMISSION

4.1. Signal Model and Synchronization Functions

In multicarrier transmission, the output of a highrate data source is split into many low-rate streams modulating adjacent subcarriers within the available bandwidth. If N is the number of the subcarriers, the symbol rate on each of them is reduced by a factor N with respect to the source rate, and this squeezes the signal bandwidth around the subcarrier to a point that the transmission channel appears to be locally flat. Correspondingly, the channel distortion on each subcarrier is reduced to a multiplicative factor that can be compensated for by a simple one-tap equalizer. The possibility of easing the equalization function has motivated the adoption of multicarrier transmission as a standard in a number of current applications, for example, European digital audiobroadcasting (DAB) and terrestrial digital videobroadcasting (DVB), IEEE 802.11 wireless local area networks, and asymmetric digital subscriber line (ADSL) and its high-speed variant VDSL, to mention only a few.

Figure 10 shows the transmitter of the most popular multicarrier transmission technique, namely, orthogonal frequency-division multiplexing (OFDM), as adopted in DAB or DVB. The input data symbols c_i at rate 1/T are serial-to-parallel (S/P)-converted and partitioned into blocks of length N. The mth block $\mathbf{c}^{(m)} = [c_0^{(m)}, c_1^{(m)}, \ldots, c_{N-1}^{(m)}]$ is fed to an N-point inverse discrete Fourier transform (IDFT) unit to produce the N-dimensional vector $\mathbf{b}^{(m)}$ (an OFDM block). The Fourier transform (FFT) techniques. Next, $\mathbf{b}^{(m)}$ is extended by appending to its end a copy of the first part of the vector, say, from $b_0^{(m)}$ to $b_{N_g-1}^{(m)}$, denoted by the cyclic prefix. The resulting extended vector drives a linear modulator with a rectangular impulse response g(t) and a signaling interval



Figure 10. OFDM transmitter.



Figure 11. OFDM receiver.

 $T_s = NT/(N + N_g)$. The cyclic prefix makes the signal insensitive to intersymbol interference, provided N_g is greater than the duration of the channel impulse response expressed in symbol intervals.

The OFDM receiver is sketched in Fig. 11. After lowpass (LP) filtering, the signal is sampled at rate $1/T_s$ and frequency and timing synchronization is performed. Next, the samples are serial-to-parallel-converted and, after removal of the cyclic prefix, they are passed to an *N*-point DFT unit whose output drives the decision device. Note that the timing circuit does not control the sampling operations. Its only purpose is to locate an appropriate window containing the samples to feed into the DFT.

Assuming perfect timing and frequency correction, the output of the DFT corresponding to the mth OFDM block is found to be

$$X^{(m)}[n] = c_n^{(m)} H[n] + w^{(m)}[n] \quad 0 \le n \le N - 1$$
(38)

where $w^{(m)}[n]$ is channel noise and H[n] is the channel response at the frequency $f_n = n/(NT)$ affecting the *n*th subcarrier. From (38) it is seen that the frequency selectivity of the channel only appears as a multiplicative term (amplitude/phase factor) on each data symbol. Accordingly, channel equalization can be performed in the frequency domain through a bank of complex multipliers [40]. This requires estimation of the channel response, which is usually performed by means of suited pilot symbols inserted in the transmitted data frame.

4.2. Frequency and Timing Estimation

Timing recovery in OFDM is significantly different from that in single-carrier systems. It just amounts to estimating where the OFDM block starts. Because of the cyclic prefix, timing offsets need not be particularly small. As long as they are shorter than the difference Δ between the length of the cyclic prefix and the CIR duration, their only effect is to generate a linear phase term across the DFT output, which can be compensated for by the channel equalizer. This is seen as follows. Denoting by εT_s the timing error (less than Δ), the DFT output is found to be

$$\begin{aligned} X^{(m)}[n] &= c_n^{(m)} H[n] e^{j 2 \pi n \varepsilon / N} + w^{(m)}[n] \\ &= c_n^{(m)} H'[n] + w^{(m)}[n] \quad 0 \le n \le N - 1 \end{aligned} \tag{39}$$

In the second equality the factor $e^{j2\pi n\varepsilon/N}$ and the channel gain H[n] have been lumped together into a single term

H'[n]. This implies on one hand that the channel estimator cannot distinguish between timing errors and channel distortions and, on the other hand, that the equalizer itself can perform fine-timing synchronization.

The main problem with OFDM systems is their sensitivity to frequency errors. It can be easily argued that the accuracy of a frequency synchronizer for OFDM has to be N times larger than with a conventional equivalent monocarrier modulation. Depending on the application, the initial frequency offset can be as large as many times the subcarrier spacing 1/(NT). If not properly compensated for, it gives rise to intercarrier interference (ICI), meaning that the *n*th DFT output $X^{(m)}[n]$ depends not only on $c_n^{(m)}$ [as indicated in (39)] but also on all the other symbols within the *m*th block.

Frequency estimation in OFDM systems can be performed jointly with timing recovery by exploiting either the redundancy introduced by the cyclic prefix [41] or pilot symbols inserted at the start of the frame [42]. The timing estimate is given by the location of the maximum of a correlation computed from the time-domain samples. The phase of the correlation is then used to estimate the frequency offset. Some feedback schemes for frequency tracking are discussed in the literature [43,44].

Only single-user OFDM systems have been considered so far. Orthogonal frequency-division multiple access (OFDMA) is a multiuser and multicarrier application that has been proposed for the uplink of wireless systems and cable TV [45] because of its robustness against multipath distortion and multiuser interference. Timing and frequency synchronization in the uplink of an OFDMA system is still an open problem that is currently under investigation.

5. FRAME SYNCHRONIZATION

In any digital stream the transmitted data have always some kind of "framing" depending on the communication system characteristics. For instance, computer data are organized into 8-bit packets (bytes), which in turn may be grouped into 512-byte or 1024-byte blocks. As a second example, some data-link layer functions (e.g., those for error control) require a specific segmentation of data. In all these cases correct detection and interpretation of data requires that the receiver be synchronized with such framing. This is the task of *frame synchronization*, which is the only instance of network synchronization that we will deal with. The most common technique to achieve frame synchronization makes use of framing markers. In essence, a *sync word or unique word* (UW) is periodically inserted into the data pattern to mark the start of a frame. The receiver knows the UW in advance and performs a search for its location in the received stream. The operation is called *frame acquisition* and is usually performed on the regenerated data. Figure 12 depicts a digital correlator for frame synchronization. The receiver continuously *correlates* the incoming bits b_k with the sync word bits u_k until a match is found. Note that b_k and u_k take on the logical values 0 or 1 and the inverted XOR gates act as comparators, meaning that their output is 1 only if the two inputs have the same value. An "in-sync" condition is declared whenever a threshold λ is crossed:

$$\sum_{i=k-N+1}^{k} b_i \circ u_{k-i} \ge \lambda \tag{40}$$

where \circ denotes the comparator operation. To avoid false locks, the transmitter is prevented from sending a data pattern equal to the UW. Stuffing additional bits in the "forbidden sequence" does this.

To reduce the probability of declaring a false lock, the UW bit pattern is designed in such a way as to exhibit a low level of "off-sync" correlation. To see this point, define the cross-correlation function of the UW as

$$R_k = \sum_{i=k}^{N-1} u_i \circ u_{i-k} \ge 0 \quad k = 0, \dots, N-1$$
 (41)

Clearly R_k equals N when k = 0 (in-sync condition) and must be much smaller than N when $k \neq 0$ (off-sync condition), otherwise a false threshold crossing might occur when searching for the correct frame alignment. Barker sequences [46] have off-sync correlation less than unity. Other good sequences are found in the textbook by Wu [47].

Transmission errors in the regenerated bits may impair frame synchronization. The performance metric in this context is the *probability of missed detection* $P_{\rm MD}$, which is defined as the probability that the digital correlator in Fig. 12 fails to detect the UW because of bit errors.



Figure 12. Digital frame synchronizer.

Denoting with P_e the probability of a bit error, the following relationship between P_{MD} and P_e is found:

$$P_{\rm MD} = \sum_{i=N-\lambda+1}^{N} {\binom{N}{i} P_e^i (1-P_e)^{N-i}}$$
(42)

For low values of P_e , this equation becomes

$$P_{\rm MD} \cong \binom{N}{N-\lambda+1} P_e^{N-\lambda+1} \tag{43}$$

and indicates that $P_{\rm MD}$ decreases exponentially as λ decreases. At first sight this would suggest taking a small threshold value. Doing so, however, might result in *false alarms*, meaning that random bit patterns might cause threshold crossings. Assuming independent and equiprobable bits, the *probability of false alarm* $P_{\rm FA}$ is found to be

$$P_{\rm FA} = \frac{1}{2^N} \sum_{i=0}^{N-\lambda} \binom{N}{i} \tag{44}$$

Since $P_{\rm FA}$ increases as λ decreases, the value of λ must be chosen as a tradeoff between contrasting requirements about $P_{\rm MD}$ and $P_{\rm FA}$. Note that $P_{\rm MD}$ and $P_{\rm FA}$ both decrease exponentially with the UW length *N*. Practical values of *N* are on the order of a few tens.

BIOGRAPHIES

Marco Luise is a full professor of telecommunications at the University of Pisa, Italy. He received M.S. and Ph.D. degrees in electronic engineering from the University of Pisa. In the past, he was a research fellow of the European Space Agency (ESA) at the European Space Research and Technology Centre (ESTEC), Noordwijk, the Netherlands, and a research scientist of the Italian National Research Council (CNR), at the Centro Studio Metodi Dispositivi Radiotrasmissioni (CSMDR), Pisa, Italy. Professor Luise cochaired four editions of the Tvrrhenian International Workshop on Digital Communications, and in 1998 was the General Chairman of the URSI Symposium ISSSE '98. He has been the Technical Chairman of the 7th International Workshop on Digital Signal Processing Techniques for Space Communications and of the Conference European Wireless 2002. As a Senior Member of the IEEE, he served as editor for Synchronization of the IEEE Transactions on Communications, and is currently editor for Communications Theory of the European Transactions on Telecommunications. His main research interests lie in the broad area of wireless communications, with particular emphasis on CDMA systems and satellite communications.

Umberto Mengali received his training in electrical engineering from the University of Pisa, Italy, where he received his degree in 1961. In 1971 he got the Libera Docenza in Telecommunications from the Italian Education Ministry and in 1975 was made a professor of electrical engineering in the Department of Information Engineering of the University of Pisa. In 1994 he was a Visiting Professor at the University of Canterbury, New Zealand, as an Erskine fellow. His research interests are in the

2484 SYNCHRONIZATION IN DIGITAL COMMUNICATION SYSTEMS

area of digital communications and communication theory, with emphasis on synchronization methods and modulation techniques. He has published over 80 journal papers and has coauthored the book *Synchronization Techniques* for Digital Receivers (Plenum Press, 1997). Professor Mengali has been an editor of the *IEEE Transactions on Communications* and of the *European Transactions on Telecommunications*. He is an IEEE fellow and is listed in *American Men and Women in Science*.

Michele Morelli was born in Pisa, Italy, in 1965. He received the Laurea (cum laude) in electrical engineering and the "Premio di Laurea SIP" from the University of Pisa, Italy, in 1991 and 1992, respectively. From 1992 to 1995 he was with the Department of Information Engineering of the University of Pisa, where he received a Ph.D. degree in electrical engineering. He is currently a research fellow at the Centro Studi Metodi e Dispositivi per Radiotrasmissioni of the Italian National Research Council (CNR) in Pisa. His interests are in digital communication theory, with emphasis on synchronization algorithms for CDMA and multicarrier systems.

BIBLIOGRAPHY

- 1. U. Mengali and A. N. D'Andrea, Synchronization Techniques for Digital Receivers, Plenum Press, New York, 1997.
- H. Meyr, M. Moeneclaey, and S. A. Fechtel, *Digital Communication Receivers*, Wiley, New York, 1998.
- 3. S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- 4. G. Asheid and H. Meyr, Cycle slips in phase-locked loops: A tutorial survey, *IEEE Trans. Commun.* 2228-2241 (Oct. 1982).
- D. C. Rife and R. R. Boorstyn, Single-tone parameter estimation from discrete-time observation, *IEEE Trans. Inform. Theory* 591-598 (Sept. 1974).
- S. A. Tretter, Estimating the frequency of a noisy sinusoid by linear regression, *IEEE Trans. Inform. Theory* 832–835 (Nov. 1985).
- S. M. Kay, A fast and accurate single frequency estimator, *IEEE Trans. Acoust. Speech Signal Process.* 1987–1990 (Dec. 1989).
- M. P. Fitz, Further results in the fast estimation of a single frequency, *IEEE Trans. Commun.* 862–864 (March 1994).
- 9. M. Luise and R. Reggiannini, Carrier frequency recovery in all-digital modems for burst-mode transmissions, *IEEE Trans. Commun.* 1169-1178 (March 1995).
- U. Mengali and M. Morelli, Data-aided frequency estimation for burst digital transmission, *IEEE Trans. Commun.* 23–25 (Jan. 1997).
- M. Morelli and U. Mengali, Feedforward frequency estimation for PSK: A tutorial review, *Eur. Trans. Telecommun.* 103-116 (March/April 1998).
- M. Morelli and U. Mengali, Carrier frequency estimation for transmission over selective channels, *IEEE Trans. Commun.* 1580–1589 (Sept. 2000).
- M. G. Hebley and D. P. Taylor, The effect of diversity on a burst-mode carrier-frequency estimator in the frequencyselective multipath channel, *IEEE Trans. Commun.* 46: 553-560 (April 1998).

- A. N. D'Andrea and U. Mengali, Noise performance of two frequency-error detectors derived from maximum likelihood estimation methods, *IEEE Trans. Commun.* 793-802 (Feb./March/April 1994).
- F. M. Gardner, Demodulator Reference Recovery Techniques Suited for Digital Implementation, European Space Agency, Final Report, ESTEC Contract 6847/86/NL/DG, Aug. 1988.
- T. Alberty and V. Hespelt, A new pattern jitter-free frequency error detector, *IEEE Trans. Commun.* 159–163 (Feb. 1989).
- K. E. Scott and E. B. Olasz, Simultaneous clock phase and frequency offset estimation, *IEEE Trans. Commun.* 2263-2270 (July 1995).
- A. J. Viterbi and A. M. Viterbi, Nonlinear estimation of PSKmodulated carrier phase with application to burst digital transmission, *IEEE Trans. Inform. Theory* 543-551 (July 1983).
- J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1989.
- M. Moeneclaey and G. de Jonghe, ML-oriented NDA carrier synchronization for general rotationally symmetric signal constellations, *IEEE Trans. Commun.* 2531–2533 (Aug. 1994).
- F. M. Gardner, *Phaselock Techniques*, 2nd ed., Wiley, New York, 1979.
- A. Polydoros, R. Raheli, and C-K. Tzou, Per survivor processing: A general approach to MLSE in uncertain environments, *IEEE Trans. Commun.* 354–364 (Feb./March/April 1995).
- C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding, *IEEE Trans. Commun.* 1261–1271 (Oct. 1996).
- L. Erup, F. M. Gardner, and R. A. Harris, Interpolation in digital modems — Part II: Implementation and performance, *IEEE Trans. Commun.* 998-1008 (June 1993).
- K. H. Mueller and M. Mueller, Timing recovery in digital synchronous data receivers, *IEEE Trans. Commun.* 516–531 (May 1976).
- F. M. Gardner, A BPSK/QPSK timing-error detector for sampled receivers, *IEEE Trans. Commun.* 423-429 (May 1986).
- 27. W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- M. Oerder and H. Meyr, Digital filter and square timing recovery, *IEEE Trans. Commun.* 605-611 (May 1988).
- 29. P. Godard, Passband timing recovery in all-digital modem receiver, *IEEE Trans. Commun.* 517–523 (May 1978).
- G. Ungerboeck, Fractional tap-spacing equalizer and its consequences for clock recovery in data modems, *IEEE Trans. Commun.* 856–864 (Aug. 1976).
- M. K. Simon, Nonlinear analysis of an absolute value type of early-late-gate bit synchronizer, *IEEE Trans. Commun. Technol.* 589-596 (Oct. 1970).
- R. De Gaudenzi, F. Giannetti, and M. Luise, Signal synchronization for direct-sequence code-division multiple-access radio modems, *Eur. Trans. Telecommun.* 73–89 (Jan./Feb. 1998).
- 33. S. E. Bensley and B. Aazhang, Maximum likelihood synchronization of a single-user for code-division multiple-access communication systems, *IEEE Trans. Commun.* 392–399 (March 1998).

- R. de Gaudenzi, M. Luise, and R. Viola, A digital chip timing recovery loop for band-limited direct-sequence spreadspectrum signals, *IEEE Trans. Commun.* COM-41(11): (Nov. 1993).
- 35. M. Moeneclaey and G. De Jonghe, Tracking performance of digital chip synchronization algorithms for bandlimited direct-sequence spread-spectrum communications, *IEE Electron. Lett.* 1147–1149 (June 1991).
- 36. S. Verdú, Multiuser Detection, Cambridge Univ. Press, 1998.
- R. A. Iltis and L. Mailaender, An adaptive multiuser detector with joint amplitude and delay estimation, *IEEE J. Select. Areas Commun.* 12: 774-785 (June 1994).
- E. G. Strom and F. Malmsten, A maximum likelihood approach for estimating DS-CDMA multipath fading channels, *IEEE J. Select. Areas Commun.* 18: 132–140 (Jan. 2000).
- V. Tripathi, A. Montravadi, and V. V. Veeravalli, Channel acquisition for wideband CDMA signals, *IEEE J. Select. Areas Commun.* 18: 1483–1494 (Aug. 2000).
- H. Sari, G. Karam, and J. Janclaude, Transmission techniques for digital terrestrial TV broadcasting, *IEEE Commun. Mag.* 36: 100–109 (Feb. 1995).
- J.-J. van de Beek, M. Sandell, and P. O. Borjesson, ML estimation of time and frequency offset in OFDM systems, *IEEE Trans. Signal Process.* 1800-1805 (July 1997).
- 42. T. M. Schmidl, and D. C. Cox, Robust frequency and timing synchronization for OFDM, *IEEE Trans. Commun.* 1613–1621 (Dec. 1997).
- F. Daffara and O. Adami, A novel carrier recovery technique for orthogonal multicarrier systems, *Eur. Trans. Telecommun.* 323–334 (July–Aug. 1996).
- M. Morelli and U. Mengali, Feedback frequency synchronization for OFDM applications, *IEEE Commun. Lett.* 28–30 (Jan. 2001).
- 45. H. Sari and G. Karam, Orthogonal frequency-division multiple access and its application to CATV networks, *Eur. Trans. Telecommun.* 507–516 (Dec. 1998).
- S. W. Golomb and R. A. Scholtz, Generalized barker sequences, *IEEE Trans. Inform. Theory* IT-11: 533–537 (Oct. 1965).
- 47. W. Wu, *Elements of Digital Satellite Communications*, Comp. Science Press, Rockville, MD, 1984.

SYNCHRONOUS OPTICAL NETWORK (SONET) AND SYNCHRONOUS DIGITAL HIERARCHY (SDH)

ROGER FREEMAN* Independent Consultant Scottsdale, Arizona

1. BACKGROUND AND INTRODUCTION

SONET and SDH are similar digital transport formats that were developed for the specific purpose of providing a reliable and versatile digital structure to take advantage of the higher bit rate capacity of optical fiber. SONET is an acronym standing for *synchronous optical network*. In a similar vein, SDH stands for *synchronous digital hierarchy*. We could say that SONET has a North American flavor, and SDH has a European flavor. This may be stretching the point, because the two systems are very similar.

The original concept in developing a digital format for high-bit-rate capacity optical systems was to have just one singular standard for worldwide application. This did not work out. The United States wanted the basic bit rate to accommodate DS3 [around 50 Mbps (megabits per second)]. The Europeans had no bit rates near this value and were opting for a starting rate around 150 Mbps. Another difference surfaced for framing alternatives. The United States perspective was based on a frame of 13-row \times 180-byte columns for the 150 Mbps rate reflecting what is now called STS-3 structure. Europe advocated a 9-row \times 270-byte column STS-3 frame to efficiently transport the E1 signal (2.048 Mbps) using 4 columns of 9 bytes, based on 32 bytes/125 μ s.

The ANSI T1X1 committee approved a final standard in August 1988, with CCITT following suit, and a global SONET/SDH standard was established. This global standard was based on a 9-row frame, wherein SONET became a subset of SDH [1].

Both SONET and SDH use basic building-block techniques. As we mentioned above, SONET starts at a lower bit rate, at 51.84 Mbps. This basic rate is called STS-1 (synchronous transport signal layer 1). Lower-rate payloads are mapped into the STS-1 format, while higher rate signals are obtained by byte interleaving N frame-aligned STS-1s to create an STS-N signal. Such a simple multiplexing approach results in no additional overhead; as a consequence the transmission rate of an STS-N signal is exactly $N \times 51.84$ Mbps where N is currently defined for the values: 1, 3, 12, 24, 48, and 192 [2].

The basic building block of SDH is the synchronous transport module level 1 (STM-1) with a bit rate of 155.52 Mbps. Lower-rate payloads are mapped into STM-1, and higher-rate signals are generated by synchronously multiplexing N STM-1 signals to form the STM-N signal. Transport overhead of an STM-N signal is N times the transport overhead of an STM-1, and the transmission rate is $N \times 155.52$ Mbps. Presently, only STM-1, STM-4, STM-16, and STM-64 have been defined by the ITU-T organization [5].

Both with SONET and SDH, the frame rate is 8000 per second, resulting in a 125-µs frame period. There is high compatibility between SONET and SDH. Because of the different basic building-block size, they differ in structure: 51.84 Mbps for SONET and 155.52 Mbps for SONET. However, if we multiply the SONET rate by

Consultants in Telecommunications. He has been writing books on various telecommunication disciplines for John Wiley & sons, Inc. since 1973. Roger has seven titles which he keeps current including *Reference Manual for Telecommunication Engineers* now in 3rd edition. His website is *www.rogerfreeman.com* and his email address is rogerf67@cox.net.

^{*} Roger Freeman took an early retirement from the Raytheon Company, Equipment Division, in 1991 where he was Principal Engineer to establish *Roger Freeman Associates*, Independent



Figure 1.1. SONET multiplexing structure. (From Fig. 3, p. 4, C. A. Siller and M. Shafi, eds., Synchronous Optical Network, Synchronous Digital Hierarchy: An Overview of Synchronous Network, IEEE Press, New York, 1996 [3].)



Figure 1.2. SDH multiplexing structure. (From Fig. 4, p. 4, Siller and Shafi [3] as Fig. 8-1.)

3, developing the STS-3 signal, we do indeed have the SDH initial bit rate of 155.52 Mbps. Figures 1.1 and 1.2 compare the multiplexing structures of each. Table 1.1 lists and compares the bit rates of each standard.

Besides differing in basic building-block bit rates, SONET and SDH differ with respect to overhead

usage. These overhead differences have been grouped into two broad categories: format definitions and usage interpretation. As a result, we have opted to segregate our descriptions of each.

We must dispel a concept that has developed by the unfortunate use of words and terminology. Some believe

SONET Optical Carrier Level OC-N	SONET Electrical Level STS-N	Equivalent SDH STM-N	Line Rate (Mbps)
OC-1	STS-1	_	51.84
OC-3	STS-3	STM-1	155.52
OC-12	STS-12	STM-4	622.08
OC-24	STS-24	_	1244.16
OC-48	STS-48	STM-16	2488.32
OC-192	STS-192	STM-64	9953.28
OC-768	STS-768	STM-256	$39,813,120^a$

Table 1.1. SONET and SDH Transmission Rates

^{*a*}On the drawing boards as of this writing [2-5].

because SONET stands for *synchronous optical network*, it will operate only on optical fiber lightguide. This is patently incorrect. Any transport medium that can provide the necessary bandwidth (measured in hertz, as one would expect), will transport the requisite SONET or SDH line rates. For example, digital loss-of-signal (LoS) microwave, using heavy bit packing modulation schemes, readily transports 622 Mbps (STS-12 or STM-4) per carrier at the higher frequencies using a 40-MHz assigned bandwidth.

The objective of this article is to provide an overview of these two standards and some of their challenging innovations that make them interesting, such as the payload pointer. Section 2 of this article deals with SONET, the North American standard, and Section 3 covers SDH. Section 4 presents a summary of the two standards in tabular form.

It should be noted that we keep custom set in this work that ITU recommendations will be identified by the CCITT or CCIR logo if that document was issued prior to January 1, 1993. If the document was issued after that date, where it originated from the Telecommunications Standardization Sector of the ITU, it will be titled ITU-T Recommendation XYZ or from the Radiocommunications Sector, ITU-R XYZ.

2. SYNCHRONOUS OPTICAL NETWORK (SONET)

2.1. Synchronous Signal Structure

SONET is based on a synchronous digital signal comprised of 8-bit octets, which are organized into a frame structure. The frame can be represented by a two-dimensional map comprising N rows and M columns, where each box so derived contains one octet (or byte). The upper lefthand corner of the rectangular map representing a frame contains an identifiable marker to tell the receiver if is the start of frame.

SONET consists of a basic, first-level structure known as STS-1, which is discussed in the following paragraphs. The definition of the first level also defines the entire hierarchy of SONET signals because higherlevel SONET signals are obtained by synchronously multiplexing the lower-level modules. When lowerlevel modules are multiplexed together, the result is denoted STS-N (STS stands for synchronous transport signal), where N is an integer. The resulting format can be converted to an OC-N (OC stands for optical carrier) or STS-N electrical signal. There is an integer multiple relationship between the rate of the basic module STS-1 and the OC-N electrical equivalent signals (i.e., the rate of an OC-N is equal to N times the rate of an STS-1). Only OC-1, OC-3, OC-12, OC-24, OC-48, and OC-192 are supported by today's SONET.

2.1.1. The Basic Building Block. The STS-1 frame is shown in Fig. 2.1. STS-1 is the basic module and building block of SONET. It is a specific sequence of 810 octets (6480 bits) that includes various overhead octets and an envelope capacity for transporting payloads.¹ STS-1 is depicted as a 90-column, 9-row structure. With a frame period of 125 μ s (i.e., 8000 frames per second). STS-1 has a bit rate of 51.840 Mbps. Consider Fig. 2.1, where the order of transmission is row-by-row, from left to right. In each octet of STS-1 the most significant bit (MSB) is transmitted first.

As illustrated in Fig. 2.1, the first three columns of the STS-1 frame contain the transport overhead. These three columns have 27 octets (i.e., 9×3), 9 of which are used for the *section overhead*, with 18 octets containing the *line*



Figure 2.1. The STS-1 frame.

 1 The several reference publications use the term *byte*, meaning, in this context, an 8-bit sequence. We prefer the term *octet*. The reason is that some argue that byte is ambiguous, having conflicting definitions.



Figure 2.2. STS-1 synchronous payload envelope (SPE).

overhead. The remaining 87 columns make up the STS-1 envelope capacity, as shown in Fig. 2.2.

The STS-1 synchronous payload envelope (SPE) occupies the STS-1 envelope capacity. The STS-1 SPE consists of 783 octets and is depicted as an 87-column \times 9-row structure. In that structure, column 1 contains 9 octets and is designated as the *STS path overhead* (POH). In the SPE, columns 30 and 59 are not used for payload but are designated *fixed-stuff* columns and are undefined. However, the values used as "stuff" in columns 30 and 59 of each STS-1 SPE will produce even parity when calculating BIP-8 of the STS-1 path BIP (bit-interleaved parity) value. The POH column and fixed stuff columns are shown in Fig. 2.3. The 756 octets in the remaining 84 columns are used for the actual STS-1 payload capacity.

The STS-1 SPE may begin anywhere in the STS-1 envelope capacity. Typically, the SPE begins in one STS-1 frame and ends in the next. This is illustrated in Fig. 2.4. However, on occasion, the SPE may be wholly contained in one frame. The *STS payload pointer* resides in the transport overhead. It designates the location of the next octet where the SPE begins. Payload pointers are described in the following paragraphs.



Figure 2.3. Path overhead (POH) and the STS-1 payload capacity within the STS-1 SPE. Note that the net payload capacity of the STS-1 is only 84 columns.



Figure 2.4. STS-1 SPE typically located in STS-1 frames. (Figure courtesy of Agilent Technologies [7].)

The STS POH is associated with each payload and is used to communicate various pieces of information from the point where the payload is mapped into the STS-1 SPE to the point where it is delivered. Among the pieces of information carried in the POH are alarm and performance data [6].

2.1.2. STS-N Frames. Figure 2.5 illustrates the structure of an STS-N frame. The frame consists of a specific sequence of $N \times 810$ octets. The STS-N frame formed by octet-interleaved STS-1 and STS-M modules (< N). The transport overhead of the associated STS SPEs are not required to be aligned because each STS-1 has a payload pointer to indicate the location of the SPE or to indicate concatenation.

2.1.3. STS Concatenation. Superrate payloads require multiple STS-1 SPEs. FDDI and some B-ISDN payloads fall into this category. Concatentation means the linking together. An STS-Nc module is formed by linking N constituent STS-1s together in a fixed-phase alignment. The superrate payload is then mapped into the resulting STS-Nc SPE for transport. Such STS-Nc SPE requires an OC-N or STS-N electrical signal. Concatenation indicators contained in the second through the Nth STS payload pointer are used to show that the STS-1s of an STS-Nc are linked together.

The are $N \times 783$ octets in an STS-Nc. Such an NTS-Nc arrangement is illustrated in Fig. 2.6 and is depicted as a $N \times 87$ column \times 9-row structure. Because of the linkage, only one set of STS POHs is required in the STS-Nc SPE. Here the STS POH always appears in the first of the N STS-1s that make up the STS-Nc [10].

Figure 2.7 shows the assignment of transport overhead of an OC-3 carrying an STS-3c SPE.

2.1.4. Structure of Virtual Tributaries (VTs). The SONET STS-1 SPE with a channel capacity of 50.11 Mbps has been designed specifically to transport a DS3 tributary signal. To accommodate sub-STS-1 rate payloads such as



Figure 2.6. STS-3c concatenated SPE. (Courtesy of Agilent Technologies [7].)

DS1, the VT structure is used. It consists of four sizes: VT1.5 (1.728 Mbps) for DS1 transport, VT2 (2.304 Mbps) for E1 transport, VT3 (3.456 Mbps) for DS1C transport, and VT6 (6.912 Mbps) for DS2 transport. The virtual tributary concept is illustrated in Fig. 2.8. The four VT configurations are illustrated in Fig. 2.9. In the 87-column \times 9-row structure of the STS-1 SPE, the VTs occupy 3, 4, 6, and 12 columns respectively [6,7].

2.2. Payload Pointer

The STS payload pointer provides a method for allowing flexible and dynamic alignment of the STS SPE within the STS envelope capacity, independent of the actual contents of the SPE. SONET, by definition, is intended to be synchronous. It derives its timing from the master network clock.

Modern digital networks must make provision for more than one master clock. Examples in the United States are the several interexchange carriers which interface with local exchange carriers (LECs), each with its own master clock. Each master clock (stratum 1) operates independently. And each of these master clocks has excellent stability (i.e., better than 1×10^{-11} per month), yet there may be some small variance in time among the clocks. Assuredly they will not be phase-aligned. Likewise, SONET must take into account loss of master clock or a segment of its timing delivery system. In this case, network switches fall back on lower-stability internal

Z0 Section growth	unspecified	unspecified	H3 pointer action byte	unspecied	unspecified	unspecified	unspecified	unspecified	1 byte (8 bits)
Z0 Section growth	unspecified	unspecified	H3 pointer action byte	unspecied	unspecified	unspecified	unspecified	unspecified	box represents
J0 Section trace	F1 user option	D3 data com channel	H3 pointer action byte	K2 automatic protection switching	D6 data channel	D9 data channel	D12 data channel	E2 express orderwire	Each
A2 framing	unspecified	unspecified	H2* pointer	unspecied	unspecified	unspecified	unspecified	M1 REI (2)	
A2 framing	unspecified	unspecified	H2* pointer	unspecied	unspecified	unspecified	unspecified	Z2 line growth	
A2 framing	E1 orderwire	D2 data channel	H2 pointer	K1 automatic protection switching	D5 data channel	D8 data channel	D11 data channel	Z2 line growth	
A1 framing	unspecified	unspecified	H1* pointer	B2 BIP-8 parity (1)	unspecified	unspecified	unspecified	Z1 line growth	
A1 framing	unspecified	nnspecified	H1* pointer	B2 BIP-8 parity (1)	unspecified	unspecified	unspecified	Z1 line growth	rleaved Parity e Error Indication
A1 framing	B1 BIP-8 parity (1)	D1 section data com channel	H1 pointer	B2 BIP-8 parity (1)	D4 data channel	D7 data channel	D10 data channel	S1 synchroni- zation	(1) BIP = Bit Inter (2) REI = Remote
	Section overhead	L	I		Line overhead	<u>.</u>	<u>.</u>		

Figure 2-7. STS-3 Transport overhead assignment. *Asterisk indicates (Based on Section 8.2, ANSI T.1.105-1995, Ref. 1).



Figure 2.8. The virtual tributary (VT) concept. (From Ref. 7, Courtesy of Agilent Technologies.)



Figure 2.9. The four sizes of virtual tributary frames. (From Ref. 7, Courtesy of Agilent Technologies.)

clocks.² The situation must be handled by SONET. Therefore, synchronous transport is required to operate effectively under these conditions, where network nodes are operating at slightly difference rates [4].

To accommodate these clock offsets, the SPE can be moved (justified) in the positive or negative direction one octet at a time with respect to the transport frame. This is accomplished by recalculating or updating the payload pointer at each SONET network node. In addition to clock offsets, updating the payload pointer also accommodates any other timing-phase adjustments required between the input SONET signals and the timing reference at the SONET node. This is what is meant by *dynamic alignment*, where the STS SPE is allowed to float within the STS envelope capacity.

The payload pointer is contained in the H1 and H2 octets in the line overhead (LOH) and designates the location of the octet where the STS SPE begins. These

 $^{^2}$ It is general practice in digital networks that switches provide timing supply for transmission facilities.



Figure 2.10. STS payload pointer (H1, H2) coding.

two octets are illustrated in Fig. 2.10. Bits 1 through 4 of the pointer word carry the *new data flag* (NDF), and bits 7 through 16 carry the pointer value. Bits 5 and 6 are undefined.

Let us discuss bits 7 through 16, the pointer value. This is a binary number with a range of 0 to 782. It indicates the offset of the pointer word and the first octet of the STS SPE (i.e., the J1 octet). The transport overhead octets are not counted in the offset. For example, a pointer value of 0 indicates that the STS SPE starts in the octet location that immediately follows the H3 octet, whereas an offset of 87 indicates that it starts immediately after the K2 octet location. These overhead octets are shown in Fig. 2.7.

Payload pointer processing introduces a signal impairment known as *payload adjustment jitter*. This impairment appears on a received tributary signal after recovery from a SPE that has been subjected to payload pointer changes. The operation of the network equipment processing the tributary signal immediately downstream is influenced by this excessive jitter. By careful design of the timing distribution for the synchronous network, payload jitter adjustments can be minimized, thus reducing the level of tributary jitter that can be accumulated through synchronous transport.

2.3. The Three Overhead Levels of SONET

The three embedded overhead levels of SONET are

- 1. Path (POH)
- 2. Line (LOH)
- 3. Section (SOH)

These overhead levels, represented as spans, are illustrated in Fig. 2.11. One important function carried out by this overhead is the support of network operation, administration, and maintenance (OA&M),

The path overhead (POH) consists of 9 octets and occupies the first column of the SPE, as pointed out previously. It is created by and included in the SPE as part of the SPE assembly process. The POH provides the facilities to support and maintain the transport of the SPE between path terminations, where the SPE is assembled and disassembled. Among the POH specific functions are

• An 8-bit wide (octet B3) BIP (bit-interleaved parity) check calculated over all bits of the previous SPE. The computed value is placed in the POH of the following frame.



Figure 2.11. SONET section, line, and path definitions.

- Alarm and performance information (octet G1).
- A path signal label (octet C2); gives details of SPE structure. It is 8 bits wide, which can identify up to 256 structures (2⁸).
- One octet (J1) repeated through 64 frames can develop an alphanumeric message associated with the path. This allows verification of continuity of connection to the source of the path signal at any receiving terminal along the path by monitoring the message string.
- An orderwire for network operator communications between path equipment (octet F2).

Facilities to support and maintain the transport of the SPE between adjacent nodes are provided by the line and section overhead. These two overhead groups share the first three columns of the STS-1 frame. The SOH occupies the top three rows (total of 9 octets, and the LOH occupies the bottom 6 rows (18 octets).

The line overhead functions include:

- Payload pointer (octets H1, H2, and H3) (each STS-1 in an STS-N frame has its own payload pointer)
- Automatic protection switching control (octets K1 and K2)
- BIP parity check (octet B2)
- 576-kbps data channel (octets D4-D12)
- Express orderwire (octet E2)

A section is defined in Fig. 2.11. Section overhead functions include [6,7]

- Frame alignment pattern (octets A1 and A2)
- STS-1 identification (octet C1): a binary number corresponding to the order of appearance in the STS-N frame, which can be used in the framing and deinterleaving process to determine the position of other signals
- BIP-8 parity check (octet B1): section error monitoring
- Data communications channel (octets D1, D2, and D3)
- Local orderwire channel (octet E1)
- User channel (octet F1)

2.4. SPE Assembly–Disassembly Process

Payload mapping is the process of assembling a tributary signal into an SPE. It is fundamental to SONET operation. The payload capacity provided for each individual tributary signal is always slightly greater than that required by that tributary signal. The mapping process, in essence, is to synchronize the tributary signal with the payload capacity. This is achieved by adding stuffing bits to the bitstream as part of the mapping process.

An example might be a DS3 tributary signal at a nominal bit rate of 44.736 of the 49.54 Mbps provided by an STS-1 SPE. The addition of path overhead completes the assembly process of the STS-1 SPE and increases the







Figure 2.13. The SPE disassembly process. (Courtesy of Agilent Technologies [7].)

bit rate of the composite signal to 50.11 Mbps. The SPE assembly process is shown graphically in Fig. 2.12. At the terminus or drop point of the network, the original DS3 payload must be recovered, as in our example. The process of SPE disassembly is shown in Fig. 2.13. The term used in this case is *payload demapping*.

The demapping process desynchronizes the tributary signal from the composite SPE signal by stripping off the path overhead and the added stuff bits. In the example, an STS-1 SPE with a mapped DS3 payload arrives at the tributary disassembly location with a signal rate of 50.11 Mbps. The stripping process results in a discontinuous signal representing the transported DS3 signal with an average signal rate of 44.74 Mbps. The timing discontinuities are reduced by means of a desynchronizing phase-locked loop, which then produces a continuous DS3 signal at the required average transmission rate of 44.736 Mbps [1,6,7].

2.5. Add/Drop Multiplex (ADM)

A SONET ADM multiplexes one or more DS-n signals into a SONET OC-N channel. In its converse function, a SONET ADM demultiplexes a SONET STS-n configuration into its component DS-n components to be passed to a user or to be forwarded on a tributary bitstream. An ADM can be configured for either the add/drop or terminal mode. In the ADM mode, it can operate when the low-speed DS1 signals terminating at the SONET derive timing from the same or equivalent source (i.e., synchronous) as the SONET system it interfaces with, but do not derive timing from asynchronous sources.

Figure 2.14 is an example of an ADM configured in the add/drop mode with DS1 and OC-N interfaces. A SONET ADM interfaces with two full-duplex OC-N signals and one or more full-duplex DS1 signals. It may optionally provide low-speed DS1C, DS2, DS3, or OC-M (where $M \leq N$).



Figure 2.14. SONET ADM add/drop configuration example [2,8].

There are non-path-terminating information payloads from each incoming OC-N signal, which are passed to the SONET ADM and transmitted by the OC-N interface on the other side.

Timing for transmitted OC-N is derived from either an external synchronization source, an incoming OC-N signal, from each incoming OC-N signals in each direction (called *through timing*), or from its local clock, depending on the network application. Each DS1 interface reads data from an incoming OC-N and inserts data into an outgoing OC-N bit stream as required. Figure 2.14 also shows a synchronization interface for local switch application with external timing and an operations interface module (OIM) that provides local technician orderwire,³ local alarm and an interface to remote operations systems. A controller is part of each SONET ADM, which maintains and controls ADM functions, to connect to local or remote technician interfaces, and to connect to required and optional operations links that permit maintenance, provisioning, and testing.

Figure 2.15 shows an example of an ADM in the terminal mode of operation with DS1 interfaces. In this case, the ADM multiplexes up to NX(28DS1) or equivalent signals into an OC-N bitstream.⁴ Timing for this terminal configuration is taken from either an external synchronization source, the received OC-N signal (called *loop timing*) or its own local clock, depending on the network application [8].

2.6. Automatic Protection Switching (APS)

First, we distinguish 1 + 1 protection from N + 1 protection. These two SONET linear APS options are shown in Fig. 2.16. APS can be provided in a linear or ring architecture. SONET NEs (network elements) that have *line termination equipment* (LTE) and terminate optical lines may provide *linear* APS. Support of linear APS at



Figure 2.15. A SONET or SDH ADM in a terminal configuration Refs. 2,8, and 12.

STS-N electrical interfaces is not provided in the relevant Telcordia and ANSI standards.

Linear APS, and in particular, the protocol for the APS channel, is standardized to allow interworking between SONET LTEs from different vendors. Therefore, all the STS SPEs carried in an OC-N signal are protected together. The ANSI and Telcordia standards define two linear APS architectures:

1+1N+1 (also called 1:n/1:1)

The 1 + 1 is an architecture in which the headend signal is continuously bridged to working and protection equipment so the same payloads are transmitted identically to the tailend working and protection equipment (see top of Fig. 2.16). At the tailend, the working and protection OC-N signals are monitored independently and identically for failures. The receiving equipment chooses either the working or protection signals as the one from which to select

³ An *orderwire* is a voice or keyboard-printer-display circuit for coordinating setup and maintenance activities among technicians and supervisors (related word: *service channel*).

⁴ This implies a DS3 configuration as it contains 28 DS1s.



Figure 2.16. (a) Linear SONET APS, 1 + 1 protection Refs. 1,9, and 12; (b) Linear SONET APS, N + 1 protection Refs. 1,9, and 12.

traffic, based on the switch initiation criteria [e.g., loss of signal (LoS), signal degraded]. Because of the continuous headend bridging, the 1 + 1 architecture does not allow an unprotected extra traffic channel to be provided.

To achieve full redundancy, 1 + 1 protection is very effective. This type of configuration is widely employed usually with a ring architecture. In the basic configuration of a ring, the traffic from the source is transmitted simultaneously over both bearers and the decision to switch between main and standby is made at the receiving location. In this situation only *loss of signal* or similar indications are required to initiate changeover and no command and control information needs to be passed between the two sites. It is assumed that after the failure in the mainline, a repair crew will restore it to service. Rather than have the repaired line placed back in service as the "mainline," it is designated the new "standby." Thus only one line interruption takes place and the process of repair does not require a second break in service.

The best method of configuring 1 + 1 service is to have the standby line geographically distant from the mainline. This minimizes common-mode failures. Because of its simplicity, this approach assures the fastest restoration of service with the least requirement for sophisticated monitoring and control equipment. However, it is more costly and involves less efficient equipment usage than does a N + 1 approach. It is inefficient in that the standby equipment sits idle nearly all the time, not bringing in any revenue.

The N + 1 (also denoted as 1:n or one for n) protection is an architecture in which any of n working lines can be bridged to a single protection line (see bottom of Fig. 2.16). Permissible values for n are 1-14. The APS channel refers to the K1 and K2 bytes in the line overhead (LOH), which are used to accomplish headend-tailend signaling. Because the headend is switchable, the protection line can be used to carry an extra traffic channel. Some texts call a subset of N + 1 architecture the 1:1 architecture.

The N + 1 link protection method makes more efficient use of standby equipment. It is merely an extension of the 1 + 1 technique described above. With the excellent reliability of present-day equipment, we can be fairly well assured that there will not be two simultaneous failures on a route. This makes it possible to share the standby line among N working lines.

Although N + 1 link protection makes more costeffective use of equipment, it requires more sophisticated control and cannot offer the same level of availability as 1 + 1 protection. Diverse routing of main and standby lines is also much harder to achieve.

2.7. SONET Ring Configurations

A ring network consists of network elements connected in a point-to-point arrangement that forms an unbroken circular configuration as shown in Fig. 2.17. As we must realize, the main reason for implementing path-switched rings is to improve network survivability. The ring provides protection against fiber cuts and equipment failures.

Various terms are used to describe path switched ring functionality, for example, *unidirectional path-protectionswitched* (UPPS) *rings*, *unidirectional path-switched rings* (UPSRs), *uniring*, *and counterrotating rings*.

Ring architectures may be considered a class of their own, but we analyze a ring conceptually in terms of 1 + 1protection. Usually, when we think of a ring architecture, we think of route diversity; and there are two separate directions of communications. The ring topology is most popular in the long-haul fiberoptic community. It offers what is called *geographic diversity*. Here we mean that there is sufficient ring diameter (e.g., >10 mi) that there is an excellent statistical probability that at least one side of the ring will survive forest fires, large floods, hurricanes, earthquakes, and other force majeure events. It also means that only one side of the ring will suffer an ordinary nominal equipment failure or "backhoe fade." Some forms of ring topology are used in CATV HFC systems, but more for achieving efficient and cost-effective connectivity than as a means of survivability. Rings are not used in building and campus fiber routings.

There are two basic SONET self-healing ring (SHR) architectures: the unidirectional and the bidirectional. Depending on the traffic demand pattern and some other factors, some ring types may be better suited to an application than others.

In a unidirectional SHR, shown in Fig. 2.17a, working traffic is carried around the ring in one direction only. For example, traffic going from node A to node D would



2. Bidirectional ring

Figure 2.17. Ring definitions—working path direction(s) Refs. 1, 8, 9, 10, 11, and 12.

traverse the ring in a clockwise direction, and traffic going from node D to node A would also traverse the ring in a clockwise direction. The capacity of a unidirectional ring is determined by the total traffic demands between any two node pairs on the ring.

In a bidirectional SHR (Fig. 2.17b), working traffic is carried around the ring in both directions, using two parallel paths between nodes (e.g., sharing the same fiber cable sheath). Using an example similar to that described above, traffic going from node A to node D would traverse the ring in a clockwise direction through intermediate nodes B and C, and traffic going from node D to node A would return along the same path also going through intermediate nodes B and C.

In a bidirectional ring, traffic in both directions of transmission between two nodes traverse the same set of nodes. Thus, unlike a unidirectional ring time slot, a bidirectional ring time slot can be reused several times on the same ring, allowing better utilization of capacity. All the nodes on the ring share the protection bit rate capacity, regardless of the number of times the time slot has been reused. Bidirectional routing is also useful on large rings, where propagation delay can be a consideration, because it provides a mechanism for ensuring that the shortest path is used (under normal conditions) against failures affecting both working and protection paths as well as node failures [6,9].

2.7.1. UPSR-BLSR Comparison. As was discussed previously, the UPSR and the BLSR have different characteristics. The UPSR offers simplicity and efficiency in hubbed traffic environments. The BLSR, although more complex than the UPSR, offers protection from some failures against which the UPSR cannot provide protection. In addition, in traffic environments that are *not* hubbed, the BLSR can potentially offer more efficiency than a UPSR. Table 2.1 compares these two techniques.

Table 2.2 contains ring capacities for different types of rings. The span capacity is the capacity of an individual span between two nodes. The ring capacity is the total bit rate capacity of all connections through a ring. Capacities are best expressed as equivalent STS-1s.

3. SYNCHRONOUS DIGITAL HIERARCHY (SDH)

3.1. Introduction

SDH resembles SONET in most respects. It uses different terminology, often for the same function as SONET. It is behind SONET in maturity by several years. History tells us that SDH will be more pervasive worldwide than SONET and is or will be employed in all countries using an E1-based PDH (plesiochronous digital hierarchy).

3.2. SDH Standard Bit Rates

SDH bit rates are built on the basic rate of STM-1 (synchronous transport module 1) of 155.520 Mbps.

UPSR (Unidirectional Path-Switched Ring)	BLSR (Bidirectional Line-Switched Ring)
Path uses bit rate capacity around entire ring	More efficient bit rate capacity utilization
Less complex, no switching protocol	Switching protocol complicated
More economic	Less economic
Used in access networks	Basic use in inter-switch (trunk) applications
Interoperability: Different node on a UPSR can be from different manufacturers	Such interoperability yet to be estabilished

Table 2.1. Comparison of SONET UPSR with SONET BLSR

Refs. 2, 10, 11, and 12.

an OC-N Ring							
Ring Type	Maximum Span Capacity	Maximum Ring Capacity					
UPSR 2-fiber	OC-N	OC-N					
BLSR 2-fiber	OC-N/2						
BLSR 4-fiber	OC-N	\geq OC-2N \leq OC-XN (see Note)					

Table 2.2. Ring Bit Rate Capacities for

Note: Depending on the traffic pattern, X = Number of ring nodes.

	SDH Level	E SO	quivalent NET Level	Bit Rate (l	kbps)	
	1	\mathbf{S}'	TS-3/OC-3	155,5	20	
	4	ST	S-12/OC-12	622,0	80	
	16	ST	S-48/OC-48	2,488,320		
	64	STS	-192/OC-192	9,953,2	80	
1 3	RSOH	_ 270 _	bytes ————	270		
_	AU pointer	'S				
9	MSOH		51M-1 pa	ауюаа	- 5 Dytes	

Table 3.1. SDH Bit Rates

Figure 3.1. STM-1 frame structure (RSOH = regenerator section overhead; MSOH = multiplex section overhead).

Higher-capacity STMs are formed at rates equivalent to N times this basic rate. STM capacities for N = 4, N = 16, and N = 64 are defined. Table 3.1 shows the bit rates presently available for SDH (Ref. G.707 3/96) with its SONET equivalents.

The basic SDH multiplexing structure is shown in Figure 1.2 [4].

3.3. Definitions

3.3.1. Synchronous Transport Module (STM). An STM is the information structure used to support section-layer connections in the SDH. It consists of information payload and section overhead (SOH) information fields organized in a block frame structure that repeats every 125 μ s. The information is suitably conditioned for serial transmission on the selected medium at a rate that is synchronized to the network. As mentioned above, a basic STM is defined at 155.520 Mbps; this is termed STM-1. Higher-capacity STMs are formed at rates equivalent to *N* times the basic rate. STM capacities are currently defined for N = 4, N = 16, and N = 64. Figure 3.1 shows the frame structure for STM-1.

The STM-1 comprises a single Administrative Unit Group (AUG) together with the section overhead (SOH). The STM-N contains N AUGs together with SOH. Figure 3.2 illustrates an STM-N.

3.3.2. Virtual Container n (VC-n). A virtual container is the information structure used to support path-layer connections in the SDH. It consists of information payload and path overhead (POH) information fields organized in a block frame structure that repeats every 125 or 500 μ s. Alignment information to identify VC-*n* frame start is provided by the server network.

Two types of virtual containers have been identified:

- Lower-Order Virtual Container n: VC-n (n = 1, 2, 3). This element contains a single container n (n = 1, 2, 3) plus the lower-order virtual container POH appropriate to that level.
- Higher-Order Virtual Container n: VC-n (n = 3, 4). This element comprises either a single container n (n = 3) or an assembly of tributary groups (TUG-2s or TUG-3s), together with virtual container POH appropriate to that level.

3.3.3. Administrative unit n (AU-n). An *administrative unit* is the information structure that provides adaptation between the higher-order path layer and the multiplex section layer. It consists of an information payload (the higher-order virtual container) and an administrative unit pointer that indicates the offset of the payload frame start relative to the multiplex section frame start.

Two administrative units are defined. The AU-4 consists of a VC-4 plus an administrative unit pointer that indicates the phase alignment of the VC-4 with respect to the STM-N frame. The AU-3 consists of a VC-3 plus an administrative unit pointer that indicates the phase alignment of the VC-3 with respect to the STM-N frame. In each case the administrative unit pointer location is fixed with respect to the STM-N frame.

One or more administrative units occupying fixed, defined positions in an STM payload are termed an *administrative unit group* (AUG). An AUG consists of an homogeneous assembly of AU-3s or an AU-4.

3.3.4. Tributary Unit N (TU-n). A tributary unit is an information structure that provides adaptation between



Figure 3.2. STM-N frame structure.

the lower-order path layer and the higher-order path layer. It consists of an information payload (the lower-order virtual container) and a tributary unit pointer that indicates the offset of the payload frame start relative to the higher-order VC frame start. The TU-n (n = 2, 2, 3) consists of a VC-n together with a tributary unit pointer.

One or more tributary units, occupying fixed defined positions in a higher order VC-n payload is termed a *tributary unit group* (TUG). TUGs are defined in such a way that mixed-capacity payloads made up of differentsize tributary units can be constructed to increase flexibility of the transport network. A TUG-2 consists of a homogeneous assembly of identical TU-1s or a TU-2. A TUG-3 consists of a homogeneous assembly of TUG-2s or a TU-3.

3.3.5. Container n (n = 1-4). A *container* is the information structure that forms the network synchronous information payload for a virtual container. For each defined virtual container, there is a corresponding container. Adaptation functions have been defined for many common network rates into a limited number of standard containers. These include all of the rates defined in CCITT Rec. G.702.

3.3.6. Pointer. A *pointer* is an indicator whose value defines the frame offset of a virtual container with respect to the frame reference of the transport entity that is supported.

3.4. Conventions

The order of transmission of information in all diagrams and figures in this section is first from left to right and then top to bottom. Within each byte (octet), the most significant bit is transmitted first. The most significant bit (bit 1) is illustrated at the left in all the diagrams, figures, and tables in this section [5].

3.5. Basic SDH Multiplexing

Figure 3.3 illustrates the relationship between various multiplexing elements, which are defined in the text below. It also shows common multiplexing structures.

Figures 3.4–3.8 illustrate examples of various signals that are multiplexed using the multiplexing elements shown in Fig. 3.3.

3.6. Administrative Units in the STM-N

The STM-N payload can support N AUGs, where each AUG may consist of one AU-4 or three AU-3s.

The VC-*n* associated with each AU-*n* does not have a fixed phase with respect to the STM-N frame. The location of the first byte of the VC-*n* is indicated by the AU-*n* pointer. The AU-*n* pointer is in a fixed location in the STM-N frame. Examples are illustrated in Figs. 3.2 and 3.4-3.9.

The AU-4 may be used to carry, via the VC-4, a number of TU-n (n = 1, 2, 3) units forming a two-stage multiplex. An example of this arrangement is illustrated in Figs. 3.8a and 3.9a. The VC-n associated with each TU-n does not have a fixed-phase relationship with respect to the start of the VC-4. The TU-n pointer is in a fixed location in the VC-4, and the location of the first byte of the VC-n is indicated by the TU-n pointer.

The AU-3 may be used to carry, via the VC-3, a number of TU-n (n = 1, 2) units forming a two-stage multiplex. An example of this arrangement is illustrated in Figs. 3.8b and 3.9b. The VC-n associated with each TU-n does not have a fixed-phase relationship with respect to the start of the VC-3. The TU-n pointer is in a fixed location in the VC-3, and the location of the first byte of the VC-n is indicated by the TU-n pointer.

3.6.1. Interconnection of STM-Ns. SDH is designed to be universal, allowing the transport of a large variety of signals including those specified in CCITT Rec. G.702. However, different structures can be used for the transport of virtual containers. The following interconnection rules will be used (Ref. G.707):

1. The rule for interconnecting two AUGs based on two different types of administrative unit—namely, AU-4 and AU-3—will be to use the AU-4 structure.



Figure 3.3. Multiplexing structure overview. *Note*: G.702 tributaries associated with containers C-*x* are shown; other signals, e.g., ATM, can also be accommodated.) (From Ref. 5, Fig. 6-1/G.707, p. 6.)



Figure 3.4. Multiplexing method directly from container 1 using AU-4. [*Note:* Unshaded areas are phase-aligned. Phase alignment between the unshaded and shaded areas is defined by the pointer (PTR) and is indicated by the arrow.] (From Ref. 5, Fig. 6-2/G.707, p. 7.)





Therefore, the AUG based on AU-3 will be demultiplexed to the VC-3 or TUG-2 level according to the type of payload, and remultiplexed within an AUG via the TUG-3/VC-4/AU-4 route. This is illustrated in Fig. 3.9a,b.

2. The rule for interconnecting VC-11s transported via different types of tributary unit—namely, TU-11 and TU-12—will be to use the TU-11 structure. This is illustrated in Fig. 3.10c. VC-11, TU-11, and TU-12 are described below.

This SDH interconnection rule does not modify the interworking rules defined in ITU-T Rec. G.802 for networks based on different PDHs and speech encoding laws.

3.6.2. Scrambling. Scrambling assures sufficient bit timing content (transitions) at the NNI to maintain synchronization and alignment. Figure 3.11 is a functional block diagram of the frame synchronous scrambler. The generating polynomial for the scrambler is $1 + X^6 + X^7$.

Figure 3.6. Multiplexing method directly from container 3 using AU-3. [*Note:* Unshaded areas are phase-aligned. Phase alignment between the unshaded and shaded areas is defined by the pointer (PTR) and is indicated by the arrow.] (From Ref. 5, Fig. 6-4/G.707, p. 9.)



Figure 3.7. Multiplexing method directly from container 4 using AU-4. [*Note:* Unshaded areas are phase-aligned. Phase alignment between the unshaded and shaded areas is defined by the pointer (PTR) and is indicated by the arrow.] (From Ref. 5, Fig. 6-5/G.707, p. 10.)



AU-*n* AU-*n* pointer + VC-*n* (see clause 8)

Figure 3.8. Administrative units in an STM-1 frame: (a) with one AU-4; (b) with three AU-3s.

3.7. Frame Structure for 51.840-Mbps Interface

Low/medium-capacity SDH transmission systems based on radio and satellite technologies that are not designed for the transmission of STM-1 signals may operate at a bit rate of 51.840 Mbps across digital sections. However, this bit rate does not represent a level of the SDH or a NNI bit rate (ITU-R Rec. G.707).

The recommended frame structure for a 51.840-Mbps signal for satellite (ITU-R Rec. S.1149) and LoS microwave application (ITU-R Rec. F.750) is shown in Fig. 3.12.

3.8. Multiplexing Methods

3.8.1. Multiplexing of Administrative Units into STM-N

3.8.1.1. Multiplexing of Administrative Unit Groups (AUGs) into an STM-N. The arrangement of N AUGs

multiplexed into an STM-N is illustrated in Fig. 3.13. The AUG is a structure of 9 rows \times 261 columns plus 9 bytes in row 4 (for the AU-*n* pointers). The STM-N consists of an SOH described below and a structure of 9 rows \times ($N \times$ 261) columns with $N \times$ 9 bytes in row 4 (for the AU-*n* pointers). The N AUGs are one single-byte-interleaved into this structure and have a fixed-phase relationship with respect to the STM-N.

3.8.1.2. Multiplexing of an AU-4 via AUG. The multiplexing arrangement of a single AU-4 via the AUG is illustrated in Fig. 3.14. The 9 bytes at the beginning of row 4 are assigned to the AU-4 pointer. The remaining 9 rows \times 261 columns are allocated to virtual container 4 (VC-4). The phase of the VC-4 is not fixed with respect to the AU-4. The location of the first byte of the VC-4 with respect to the AU-4 pointer is given by the pointer value. The AU-4 is placed directly in the AUG.

3.8.1.3. Multiplexing of AU-3s via AUG. The multiplexing arrangement of three AU-3s via the AUG is shown in Fig. 3.15. The 3 bytes at the beginning of row 4 are assigned to the AU-3 pointer. The remaining 9 rows \times 87 columns are allocated to the VC-3 and two columns of fixed stuff. The byte in each row of the two columns of fixed stuff of each AU-3 shall be the same. The phase of the VC-3 and the two columns of fixed stuff is not fixed with respect to the AU-3. The location of the first byte of the VC-3 with respect to the AU-3 pointer is given by the pointer value. The three AU-3s are single-byte-interleaved in the AUG.



TU-n TU-n pointer + VC-n

Figure 3.9. Two-stage multiplex: STM-1 with (a) one AU-4 containing TUs and (b) three AU-3s containing TUs. (From Ref. 5, Fig. 6-8/G.707, p. 12.)



Figure 3.10. Interconnection of STM-Ns: (a) of VC-3 with C-3 payload; (b) of TUG-2; (c) of VC-11. (From Ref. 5, Fig. 6-9/G.707, p. 16.)



Figure 3.11. Functional block diagram of a frame-synchronous scrambler. (From Ref. 5, Fig. 6-10/G.707, p. 17, ITU-T.)



NOTES

1 M1 position is not the same position [9, 3N + 3] as in a STM-N frame. 2 Fixed stuff columns are not part of the VC-3.

Figure 3.12. Frame structure for 51.840-Mbps (SDH) operation. (From Ref. 5, Fig. A.1/G.707, p. 89.)



Figure 3.13. Multiplexing of N AUGs into an STM-N. (From Ref. 5, Fig. 7-1/G.707, p. 18.)

3.8.2. Multiplexing of Tributary Units into VC-4 and VC-3

3.8.2.1. Multiplexing of Tributary Unit Group 3S (TUG-3s) into a VC-4

The arrangement of three TUGs multiplexed in the VC-4 is illustrated in Fig. 3.16 The TUG-3 is a 9-row \times 86-column structure. The VC-4 consists of one column of VC-4 POH, two columns of fixed stuff and a 258-column payload structure. The three TUG-3s are single-byte-interleaved into the 9-row \times 258-column VC-4 payload structure and have a fixed phase with respect to the VC-4. As described in 3.9.1.1, the phase of the VC-4 with respect to the AU-4 is given by the AU-4 pointer.

3.8.2.2. Multiplexing of a TU-3 via a TUG-3. The multiplexing of a single TU-3 via the TUG-3 is shown in Fig. 3.17. The TU-3 consists of the VC-3 with a 9-byte VC-3 POH and the TU-3 pointer. The first column of the 9-row \times 86-column TUG-3 is assigned to the TU-3 pointer (bytes H1, H2, H3) and fixed stuff. The phase of the VC-3 with respect to the TUG-3 is indicated by the TU-3 pointer.

3.8.2.3. Multiplexing of TUG-2s via a TUG-3. The multiplexing format for the TUG-2 via the TUG-3 is illustrated in Fig. 3.18. The TUG-3 is a 9-row \times 86-column structure with the first two columns of fixed stuff.

3.8.2.4. Multiplexing of TUG-2s into a VC-3. The multiplexing structure for TUG-2s into a VC-3 is shown



Y 1001 SS11 (S bits are unspecified)

Figure 3.14. Multiplexing of AU-4 via AUG. (From Ref. 5, Fig. 7-2/G.707, p. 19.)

in Fig. 3.19. The VC-3 consists of VC-3 POH and a 9-row \times 84-column payload structure. A group of seven TUG-2s can be multiplexed into the VC-3.

3.9. Pointers

3.9.1. AU-*n* Pointer. The AU-*n* pointer provides a method of allowing flexible and dynamic alignment of the VC-*n* within the AU-*n* frame. Dynamic alignment means that the VC-*n* is allowed to "float" within the AU-*n* frame. Thus, the pointer is able to accommodate differences, not only in the phases of the VC-*n* and the SOH, but also in the frame rates.

3.9.1.1. AU-n Pointer Location. The AU-4 pointer is contained in bytes H1, H2, and H3 as shown in Fig. 3.20. The three individual AU-3 pointers are contained in three separate H1, H2, and H3 bytes as shown in Fig. 3.21.

3.9.1.2. AU-n Pointer Value. The pointer contained in H1 and H2 designates the location of the byte where the VC-*n* begins. The 2 bytes allocated to the pointer function should be viewed as one word, as illustrated in Fig. 3.22. The last ten bits (bits 7-16) of the pointer word carry the pointer value.

As illustrated in Fig. 3.22, the AU-4 pointer value is a binary number with a range of 0-782 that indicates

the offset in 3-byte increments, between the pointer and the first byte of the VC-4 (see Fig. 3.20). Figure 3.22 also indicates one additional valid pointer, the Concatenation Indication. The Concatenation Indication is indicated by binary 1001 in bits 1-4, bits 5-6 are unspecified, and 10 binary 1s in bit positions 7-16. The AU-4 pointer is set to "concatenation indication" for AU-4 concatenation.

As shown in Fig. 3.22, the AU-3 pointer value is also a binary number with a range of 0-782. Since there are three AU-3s in the AUG, each AU-3 has its own associated H1, H2, and H3 bytes.

Note that the H bytes are shown in sequence in Fig. 3.20. The first H1,H2,H3 set refers to the first AU-3, and the second set to the second AU-3, and so on. For the AU-3s, each pointer operates independently.

In all cases, the AU-n pointer bytes are not counted in the offset. For example, in an AU-4, the pointer value of 0 indicates that the VC-4 starts in the byte location that immediately follows the last H3 byte, whereas an offset of 87 indicates that the VC-4 starts 3 bytes after the K2 byte.

3.9.1.3. Frequency Justification. If there is a frequency offset between the frame rate of the AUG and that of the VC-*n*, then the pointer value will be incremented or decremented as needed, accompanied by a corresponding positive or negative justification byte or bytes. Consecutive pointer operations must be separated by at least three



Figure 3.15. Multiplexing of AU-3s via AUG. (*Note*: The byte in each row of the two columns of fixed stuff of each AU-3 shall be the same.) (From Ref. 5, Fig. 7-3/G.707, p. 20.)



Figure 3.16. Multiplexing of three TUG-3s into a VC-4. (From Ref. 5, Fig. 7/4/G.707, p. 21.)

frames (i.e., every fourth frame) in which the pointer value remains constant.

If the frame rate of the VC-n is too slow with respect to that of the AUG, then the alignment of the VC-n must periodically slip back in time and the pointer value must be incremented by one. This operation is indicated by inverting bits 7, 9, 11, 13, and 15 (I bits) of the pointer word to allow 5-bit majority voting at the receiver. Three positive justification bytes appear immediately after the last H3 byte in the AU-4 frame containing inverted I bits. Subsequent pointers will contain the new offset. This is depicted in Fig. 3.23.

For AU-3 frames, a positive justification byte appears immediately after the individual H3 byte of the AU-3



Figure 3.17. Multiplexing a TU-3 via a TUG-3. (From Ref. 5, Fig. 7-5/G.707, p. 21.)



Figure 3.18. Multiplexing of TUG-2s via a TUG-3. (From Ref. 5, Fig. 7-6/G.707, p. 22.)

frame containing inverted ${\cal I}$ bits. Subsequent pointers will contain the new offset.

If the frame rate of the VC-n is too fast with respect to that of the AUG, then the alignment of the VC-n must periodically be advanced in time and the pointer value must be decremented by one. This operation is indicated by inverting bits 8, 10, 12, 14, and 16 (D bits) of the pointer word to allow 5-bit majority voting at the receiver. Three negative justification bytes appear in the H3 bytes in the AU-4 frame containing inverted D bits. Subsequent pointers will contain the new offset.

For AU-3 frames, a negative justification byte appears in the individual H3 byte of the AU-3 frame containing inverted D bits. Subsequent pointers will contain the new offset. **3.9.1.4. Pointer Generation.** The following summarizes the rules for generating the AU-*n* pointers:

- 1. During normal operation, the pointer locates the start of the VC-*n* within the AU-*n* frame. The NDF (new data flag) is set to binary 0110. The NDF consists of the *N* bits, bits 1-4 of the pointer word.
- 2. The pointer value can be changed only by operation 3, 4, or 5 (see items below).
- 3. If a positive justification is required, the current pointer value is sent with the I bits inverted and the subsequent positive justification opportunity is filled with dummy information. Subsequent pointers contain the previous pointer value incremented by one. If the previous pointer is at its maximum value,



Figure 3.19. Multiplexing seven TUG-2s into a VC-3. (From Ref. 5, Fig. 7-8/G.707, p. 24.)



^{1*} All 1s byte

the subsequent pointer is set to zero. No subsequent increment or decrement operation is allowed for at least three frames following this operation.

4. If a negative justification is required, the current pointer value is sent with the D bits inverted and the subsequent negative justification opportunity is overwritten with actual data. Subsequent pointers contain the previous pointer value decremented by one. If the previous value is zero, the subsequent



pointer is set to its maximum value. No subsequent increment or decrement operation is allowed for at least three frames following this operation.

5. If the alignment of the VC-n changes for any reason other than rules 3 or 4, the new pointer value should be sent accompanied by the NDF set to 1001. The NDF only appears in the first frame that contains the new values. The new location of the VC-n begins at the first occurrence of the offset

Y 1001SS11 (S bits are unspecified)



Figure 3.22. AU-n/TU-3 pointer (H1, H2, H3) coding. (From Ref. 5, Fig. 8-3/G.707, p. 36.)

indicated by new pointer. No subsequent increment or decrement operation is allowed for at least three frames following this operation.

3.9.1.5. Pointer Interpretation. The following list summarizes the rules for interpreting the AU-*n* pointers:

- 1. During normal operation, the pointer locates the start of the VC-*n* within the AU-*n* frame.
- Any variation from the current pointer value is ignored unless a consistent new value is received 3 times consecutively or is preceded by either rule 3, 4, or 5. Any consistent new value received three times consecutively overrides (i.e., takes priority over) rules 3 and 4.
- 3. If the majority of the I bits of the pointer word is inverted, a positive justification operation



Y 1001SS11 (S bits are unspecified)

Figure 3.23. AU-4 pointer adjustment operation, positive justification. (From Ref. 5, Fig. 8-4/G.707, p. 37.)

Fable 4.1. Summary of SDH/SONET Payloads and Mappings ^a								
	SDH				SONET			
Payload	Container	Actual Payload Capacity	Payload and POH	Mapping AU-3/AU-4 Based	Container SPE	Actual Payload Capacity	Payload and POH	
DS1 (1.544)	VC-11	1.648	1.728	(AU-3),AU-4	VT 1.5	1.648	1.664	
	VC-12	2.224	2.304	AU-3,AU-4				
E1 (2.048)	VC-12	2.224	2.304	(AU-3),AU-4	VT2	2.224	2.240	
DS1C (3.152)					VT3	3.376	3.392	
DS2 (6.312)	VC-2	6.832	6.912	(AU-3),AU-4	VT6	6.832	6.848	
E3 (34.368)	VC-3	48.384	48.960	AU-3,AU-4				
DS3 (44.736)	VC-3	48.384	48.960	(AU-3),AU-4	STS-1	49.536	50.112	
E4 (139.264)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	
ATM (149.760)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	
ATM (599.040)	VC-4-4c	599.040	601.344	(AU-4)	STS-12c	599.040	601.344	
FDDI (125.000)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	
DQDB (149.760)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	

 $^{a}(AU-n)$ indicates compatible mapping to SONET Note 2: Numbers are in Mbit/s unit.

Source: Shafi and Siller, Ref. 3 Table 2, p. 64, 3-, reprinted with permission.

is indicated. Subsequent pointer values shall be incremented by one.

- 4. If the majority of the D bits of the pointer word is inverted, a negative justification operation is indicated. Subsequent pointer values shall be decremented by one.
- 5. If the NDF is interpreted as enabled, the coincident pointer value shall replace the current one at the

offset indicated by the new pointer value unless the receiver is in a state that corresponds to a loss of pointer.

Mapping STS-1 STS-1 STS-1 STS-1

STS-1 STS-3c STS-3c STS-12c STS-3c

STS-3c

4. SONET/SDH SUMMARY

Table 4.1 summarizes SDH/SONET payloads and mappings. Table 4.2 reviews the various overheads for SDH STM-1 and SONET STS-3c.
				Path	Overhead					_
Trace J1	BIP-8 B3	Sig Label C2	Path status G1	User F2	Multiframe H4	Growth Z3	Growth Z4	Growth Z5	ıyload capacity	
	For STS-30	only, not included in STM-1							Å.	
STS-1 ID C1			Ptr Action H3							_
STS-1 ID C1			Ptr Action H3							
STS-1 ID C1	User F1	Data Communication D3	Ptr Action H3	APS K2	Data Communication D6	Data Communication D9	Data Communication D12	Orderwire E2		
Framing A2			1111 1111					Growth Z2	ad	
Framing A2			1111 1111					Growth Z2	port overhea	
Framing A2	Order- wire E1	Data Communication D2	Pointer H2	APS K1	Data Communication D5	Data Communication D8	Data Communication D11	Growth Z2	Trans	inted with normiseion
Framing A1			$\frac{1001}{\mathrm{ss}11}$	BIP-8 B2				Growth Z1		vuon 64 mores
Framing A1			$\frac{1001}{\mathrm{ss}11}$	BIP-8 B2				Growth Z1		ն ըչք Չ ՄշԽՍ
Framing A1	BIP-8 B1	Data Communication D1	Pointer H1	BIP-8 B2	Data Communication D4	Data Communication D7	Data Communication D10	Growth Z1		Correct Cillor and Cha

Table 4.2. Overhead Summary, STM-1, STS-3c

2509

BIOGRAPHY

Roger Freeman has over 50 years experience in telecommunications including a stint in the US Navy and radio officer on merchant vessels. He attended Middlebury College and has two degrees from New York University. He has had assignments with the Bendix Corporation in Spain and North Africa which was followed by five years as a member technical staff for ITT Communications Systems. Roger then became manager of microwave systems for CATV extension at Jerrold Electronics Corporation followed by assignments at Page Communications Engineers in Washington, DC where he was a project engineer on earth stations and on various data communication programs. During this period he was assigned by the ITU as Regional Planning Expert for northern Sourth America based in Quito, Ecuador. From Quito he took a position with ITT at their subsidiary in Madrid, Spain where he did consulting in telecommunication planning. In 1978 he joined the Raytheon Company as principal engineer in their Communication Systems Directorate where he held design positions on military communications. At the same time he taught various telecommunication courses in the evenings at Northeastern University and 4-day seminars at the University of Wisconsin. These seminars were based on his several textbooks on telecommunications published by John Wiley & Sons, New York. He also gives telecommunication seminars (in Spanish) in Monterrey, Mexico City and Caracas. Roger is a contributor and guest editor (Desert Storm edition) of the IEEE Communications magazine and was advanced by the IEEE to senior life member in 1994. He served on the board of directors of the Spain Section of the IEEE and was its secretary for four years. In 1991 Roger took early retirement from the Raytheon Company and organized Roger Freeman Associates, Independent Consultants in Telecommunications. The group has undertaken over 50 assignments from Alaska to South America.

Roger may be reached at *rogerf67@cox.net*; his website is *www.rogerfreeman.com*. Also of interest would be *www.telecommunicationbooks.com* where the reader may subscribe to the on-line Reference Manual for Telecommunication Engineering, 3rd ed, updated quarterly.

BIBLIOGRAPHY

- Synchronous Optical Network (SONET)—Basic Description Including Multiplex Structure, Rates and Formats, ANSI T1.105-1995, ANSI New York, 1995.
- 2. R. L. Freeman, Reference Manual for Telecommunications Engineering, 3rd ed., Wiley, New York, 2001.
- C. A. Siller and M. Shafi, eds., SONET/SDH, IEEE Press, New York, 1996.
- 4. R. L. Freeman, *Telecommunication Transmission Handbook*, 4th ed., Wiley, New York, 1998.
- Network-Node Interface for the Synchronous Digital Hierarchy (SDH), ITU-T Rec. G.707, ITU Geneva, March, 1996.
- Synchronous Optical Network (SONET), Transport Systems, Common Generic Criteria, Telcordia GR-253-CORE, Issue 2, Rev. 2, Piscataway, NJ, Jan. 1999.
- Introduction to SONET, Seminar, Hewlett-Packard, Burlington, MA, Nov. 1993.
- SONET Add-Drop Multiplex Equipment (SONET ADM) Generic Criteria, Bellcore, TR-TSY-000496, Issue 2, Bellcore, Piscataway, NJ, 1989.
- 9. Automatic Protection Switching for SONET, Telcordia Special Report SR-NWT-001756, Issue 1, Telcordia, Piscataway, NJ, Oct. 1990.
- SONET Dual-Fed Unidirectional Path Switched Ring (UPSR) Equipment Generic Criteria, Telcordia GR-1400-CORE, Issue 2, Telcordia, Piscataway, NJ, Jan. 1999.
- SONET Bidirectional Line-Switched Ring Equipment Generic Criteria, Telcordia GR-1230-CORE, Issue 4, Telcordia, Piscataway, NJ, Dec. 1998.
- Telcordia Notes on the Synchronous Optical Network (SONET), Special Report, SR-NOTES-Series-01, Issue 1, Piscataway, NJ, Dec. 1999.

TAILBITING CONVOLUTIONAL CODES

Marc Handlery Rolf Johannesson Per Ståhl Lund University Lund, Sweden

1. INTRODUCTION

Error-correcting codes should protect digital data against errors that occur on noisy communication channels. There are two basic types of error-correcting codes: block codes and convolutional codes, which are similar in some ways but differ in many other ways. For simplicity, we consider only binary codes.

We first consider block codes and divide the entire sequence of data or information digits into blocks of length K called *information words*. Then the *block encoder* maps the set of information words one to one to the set of *codewords* of *blocklength* N, where $N \ge K$. The block code \mathcal{B} is the set of $M = 2^{K}$ codewords and the block code *rate* R = K/N is the fraction of binary digits in the codeword that is needed to represent the information word; the remaining fraction, 1 - R, represents the *redundancy* that can be used to combat channel noise. The theory of block codes is very rich, and many algebraic concepts have been used to design block codes with mathematical structures that can be exploited in efficient decoding algorithms. Block codes are used in many applications; for example, all compact-disk (CD) players use a very powerful block code called the *Reed-Solomon code*.

The encoder for a convolutional code maps the sets of information words (in this context also called an *information sequences*) one to one to the set of codewords. An information word is regarded as a continuous sequence of *b*-tuples, and a codeword is regarded as a continuous sequence of *c*-tuples. The convolutional encoder introduces dependencies between the codeword *c*-tuples; the present codeword *c*-tuple depends not only on the corresponding information word *b*-tuple but also on the *m* (memory) previous *b*-tuples. The *convolutional code* is the (infinite) set of infinitely long codewords and the convolutional code rate is R = b/c. For block codes, *K* and *N* are usually large, whereas for convolutional codes, *b* and *c* are usually small. Convolutional codes are routinely used in many applications, for example, mobile telephony and modems.

A major drawback of many block codes is that it is difficult to make use of soft-decision information provided by a channel (or demodulator) that, instead of outputting a hard decision of a channel symbol, outputs likelihood information. Such channels are called *softoutput channels* and are often better models of the situation encountered in practice. For convolutional codes there exist decoding algorithms that can easily exploit the soft-decision information provided by a soft-output channel. This is one reason why convolutional codes are often used in practice.

Tailbiting convolutional codes, in the sequel simply called *tailbiting codes*, are block codes that are obtained from convolutional codes and can be regarded as a link between these two types of codes. They inherit many properties, such as distance properties and the error-correcting capability, from convolutional codes. Many decoding algorithms for convolutional codes that make use of softdecision information can be extended to tailbiting codes.

The basic idea of obtaining good block codes from convolutional codes by using the tailbiting method was first presented by Solomon and van Tilborg [1]. The term *tailbiting convolutional code* was, however, not used until Ma and Wolf presented their results in 1986 [2]. It has since been shown that many good block codes can be considered as tailbiting codes. Tailbiting codes are also an interesting option in practical applications where information is to be transmitted in rather large blocks and where the advantages of convolutional codes can be exploited.

2. TERMINATING CONVOLUTIONAL CODES

In order to explain the tailbiting termination method, a short introduction to convolutional codes and convolutional encoders is first given. A rate R = b/c, binary convolutional encoder has *b* inputs and *c* outputs and encodes a semiinfinite information sequence $\mathbf{u} = \mathbf{u}_k \mathbf{u}_{k+1} \mathbf{u}_{k+2} \dots$, where $\mathbf{u}_i = (u_i^{(1)} \ u_i^{(2)} \cdots u_i^{(b)})$ is a binary *b*-tuple and *k* is an integer (positive or negative), into a code sequence $\mathbf{v} = \mathbf{v}_k \mathbf{v}_{k+1} \mathbf{v}_{k+2} \dots$, where $\mathbf{v}_i = (v_i^{(1)} \ v_i^{(2)} \dots v_i^{(c)})$ is a binary *c*-tuple. For simplicity we will first only consider convolutional encoders without feedback. In this case the code symbol \mathbf{v}_t at time *t* depends on both the information symbol \mathbf{u}_t and the *m* previous information symbols. The encoding rule can be written as

$$\mathbf{v}_t = \mathbf{u}_t G_0 + \mathbf{u}_{t-1} G_1 + \dots + \mathbf{u}_{t-m} G_m \tag{1}$$

where $t \ge k$ is an integer and $\mathbf{u}_j = 0$ when j < k. The parameter *m* is called the *memory* of the encoder and G_i , $0 \le i \le m$, is a binary $b \times c$ matrix. The arithmetic in (1) is in the binary field \mathbb{F}_2 . A convolutional code encoded by a convolutional encoder is the set of all code sequences **v** obtained from the encoder resulting from all possible different information sequences **u**. For simplicity we will assume here that k = 0.

The convolutional encoder can be regarded as a linear finite-state machine. Figure 1 shows an example of a code rate $R = \frac{1}{2}$ convolutional encoder. It has the following encoding rule

$$\mathbf{v}_t = \mathbf{u}_t(1\ 1) + \mathbf{u}_{t-1}(1\ 0) + \mathbf{u}_{t-2}(1\ 1)$$
(2)

where \mathbf{u}_j , j = t, t - 1, t - 2, is a binary 1-tuple, and its memory is m = 2. The encoder has two delay elements;

hence, at each time t the encoder can be in four different states depending on the two previous information symbols. We denote the state at time t by \mathbf{s}_t , the set of all possible encoder states by S, and the number of possible encoder states by n_s . A convolutional encoder is assumed to start in the all-zero state: $\mathbf{s}_0 = \mathbf{0}$. A trellis [3] describes all possible state sequences $\mathbf{s}_0 \mathbf{s}_1 \mathbf{s}_2 \cdots$ of the convolutional encoder. The trellis for the rate $R = \frac{1}{2}$ memory m = 2 encoder in Fig. 1 is shown in Fig. 2. From each state in the trellis there are at each time instant two possible state transitions. The state transitions are represented by branches. The upper branch corresponds to the encoder input 0 and the lower one, to the encoder input 1. The symbols on each branch denote the encoder output.

A convolutional code is a set of (semi)infinitely long code sequences resulting from (semi)infinite strings of data. However, since data usually are transmitted in packets of a certain size and not as a stream of symbols, it is in a practical situation necessary to split the infinite datastream into blocks, and let each block of data be separately encoded into a block of code symbols by the convolutional encoder. We say that the convolutional code is *terminated* into a block code. An encoder for an (N, K) block code maps blocks of K information bits $\mathbf{u} =$ $(u_0 \ u_1 \cdots u_{K-1})$ one to one to codewords $\mathbf{v} = (v_0 \ v_1 \cdots v_{N-1})$ of N bits.

The straightforward method for terminating a convolutional code into a block code is to use the termination method called *direct truncation*. Assume that a rate R = b/c convolutional encoder with memory m is used to encode a block of K = Lb information bits $\mathbf{u} = (\mathbf{u}_0 \ \mathbf{u}_1 \cdots \mathbf{u}_{L-1})$, where $\mathbf{u}_i = (u_i^{(1)} \ u_i^{(2)} \cdots u_i^{(b)})$. Using direct truncation, the convolutional encoder is started in a fixed state, and is simply fed with the L information b-tuples. The resulting output is a codeword $\mathbf{v} = (\mathbf{v}_0 \ \mathbf{v}_1 \ \dots \ \mathbf{v}_{L-1})$, where $\mathbf{v}_i = (v_i^{(1)} \ v_i^{(2)} \cdots v_i^{(c)})$, consisting of N = Lc bits and the set



Figure 1. A rate $R = \frac{1}{2}$ memory m = 2 convolutional encoder.



Figure 2. The trellis for the encoder in Fig. 1.

of codewords resulting from all the $M = 2^{K}$ possible different information blocks makes up a linear block code. The blocklength of this code is N = Lc, and the code's rate is $R_{\rm dt} = K/N = b/c$.

The disadvantage with the direct-truncation method is that the last information *b*-tuples are less protected than the other data bits. A remedy for this is to feed the encoder with *m* additional *b*-tuples, after the information symbols, which forces the encoder in to an advanced determined encoder state (known to the decoder). These *mb* symbols carry no information and are called "dummy symbols." Usually the predetermined ending state is chosen to be the all-zero state, and we reach this state by feeding the (feedback-free) encoder by m dummy zero b-tuples. Hence, this second termination method is called the zerotail method. With this method also the last information symbols are protected, and this is the most common method used to terminate a convolutional code. The block code obtained with the zero-tail method has blocklength N = (L + m)c and rate

$$R_{\rm zt} = \frac{K}{N} = \frac{L}{L+m} \frac{b}{c} = \frac{L}{L+m} R \tag{3}$$

which is less than the rate R = b/c of the convolutional encoder. The rate loss L/(L + m), caused by the *m* dummy *b*-tuples, is negligible if $L \gg m$, but if *L* is small, that is, if the data arrives in short packets, this rate loss might not be acceptable.

In the third method, called *tailbiting*, this rate loss is removed. In order to protect all data bits equally, we impose the restriction that the rate R = b/c convolutional encoder should start and, after feeding it with the K = Lbinformation bits **u**, end in the same encoder state. To avoid the use of dummy symbols, this starting-ending state is *not* fixed, but depends on the actual information sequence to be encoded. Since no dummy symbols are used, the codewords have length N = Lc, and the rate of the block code obtained is the same as for the encoder:

$$R_{\rm tb} = \frac{K}{N} = \frac{b}{c} = R \tag{4}$$

We have no rate loss. The following example illustrates our three different termination methods.

Example 1. Assume that the rate $R = \frac{1}{2}$ encoder with memory m = 2 in Fig. 1 is used to encode a block of 4 information bits. Figure 3a-c shows the trellis when direct truncation, zero-tail, and tailbiting termination types are used, respectively. With direct truncation the encoder starts in the all-zero state and can end in any of the four encoder states, whereas with zero-tail termination the encoder can start and end in one predetermined state, that is, in the all-zero state only. With tailbiting the encoder, as with zero-tail termination, also ends in the same state as it started in, however, all four encoder states are possible starting-ending states.

Consider the encoding of the information block $\mathbf{u} = (1\ 0\ 0\ 1)$ using the tailbiting technique. If the encoder is loaded with (starts in) state 10, then after feeding it with the information sequence \mathbf{u} , the encoder again ends in



state 10; that is, the tailbiting restriction is fulfilled. The corresponding codeword is $\mathbf{v} = (01\ 01\ 11\ 11)$. The encoder state sequence is shown in Fig. 3d.

We call a linear block code obtained by terminating a convolutional code using the tailbiting termination method a tailbiting code, and the number of encoded b-tuples L is called the *tailbiting length*. The corresponding trellis is called a *tailbiting trellis*. Since the convolutional encoder starts and ends in the same state, we can view the tailbiting trellis as a circular trellis. Figure 4 shows a circular trellis of length L = 6 for a tailbiting code encoded by a rate R = 1/c encoder with four states. Every valid codeword in a tailbiting code corresponds to a circular path in the circular trellis. From the regular structure of the circular trellis obtained from a time-invariant convolutional encoder, it follows that if a codeword is cyclically shifted c symbols (corresponding to one cyclic step in the circular trellis), the result is also a codeword in the tailbiting code. Block codes with this property are called *quasicyclic* codes. Hence, all R = b/c tailbiting codes obtained from time-invariant convolutional encoders are quasicyclic under a cyclic shift of *c* symbols.

3. ENCODING TAILBITING CODES

Using the tailbiting termination method we must find the correct starting state for each information word to be encoded such that the encoder starts and ends in Figure 3. Trellises for encoding a block of 4 information bits using the encoder in Fig. 1 and (a) direct truncation, (b) zero-tail termination, (c) tailbiting termination. In (d) the encoder state sequence when encoding $\mathbf{u} = (1 \ 0 \ 0 \ 1)$ using tailbiting is shown.



Figure 4. A circular trellis for a tailbiting code of length L = 6 encoded by a rate R = 1/c encoder with four states.

the same state. For feedback-free encoders realized in controller canonical form, like the encoder in Fig. 1, this is very easy. Consider the encoding of the information word $\mathbf{u} = (1\ 0\ 0\ 1)$ in Example 1. After feeding the convolutional encoder with the 4 information bits, it ends in state (1 0), which is simply the 2 last information bits in reversed order, and, hence, if this state is chosen as starting state, then the tailbiting criterion will be fulfilled. Similarly, for feedback-free encoders realized in controller canonical form the encoder starting state is the reverse of the *m* last

2514 TAILBITING CONVOLUTIONAL CODES

information *b*-tuples $\mathbf{u}_{L-1} \cdots \mathbf{u}_{L-m}$. If the convolutional encoder has feedback, the starting state does not depends not only on the *m* last information *b*-tuples but, in general, also on all information bits. The starting state can be obtained by performing a simple matrix multiplication as shown in Ref. 4, where finding the starting state of a convolutional encoder realized in observer canonical form is also investigated. Another method to find the starting state of a feedback convolutional encoder when tailbiting is used is given by Weiss et al. [5]. This method requires that the encoding is performed twice: once in order to find the correct starting state and a second time for the actual encoding.

From Eq. (1) and the fact that the starting state is given by the reverse of the last m information b-tuples to be encoded, it follows that for feedback-free encoders

$$\mathbf{v}_{t} = (v_{t}^{(1)} \ v_{t}^{(2)} \cdots v_{t}^{(c)}) = \sum_{k=0}^{m} \mathbf{u}_{((t-k))} G_{k}$$
(5)

where the double parentheses denote modulo *L* arithmetic on the indices; that is, $((t - k)) \equiv t - k \pmod{L}$. Then the encoding of a tailbiting code using a feedback-free encoder can be compactly written as

$$\mathbf{v} = \mathbf{u}\mathbf{G}^{\mathrm{tb}} \tag{6}$$

where

$$\mathbf{G}^{\text{tb}} = \begin{pmatrix} G_0 & G_1 & G_2 & \dots & G_m & & \\ & G_0 & G_1 & G_2 & \dots & G_m & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & G_0 & G_1 & G_2 & \dots & G_m \\ G_m & & & & G_0 & G_1 & \dots & G_{m-1} \\ G_{m-1} & G_m & & & & G_0 & \dots & G_{m-2} \\ \vdots & & \ddots & & & \ddots & \vdots \\ G_1 & G_2 & \dots & G_m & & & & G_0 \end{pmatrix}$$
(7)

is an $L \times L$ matrix and where each entry is a $b \times c$ matrix.

Example 2. The generator matrix that corresponds to the tailbiting encoding in Example 1 is

$$\mathbf{G}^{\text{tb}} = \begin{pmatrix} 11 & 10 & 11 & 00\\ 00 & 11 & 10 & 11\\ 11 & 00 & 11 & 10\\ 10 & 11 & 00 & 11 \end{pmatrix}.$$
 (8)

We verify that when encoding $u=(1\ 0\ 0\ 1)$ the corresponding codeword is $v=(1\ 0\ 0\ 1)G^{tb}=(01\ 01\ 11\ 11).$

Often it is convenient to express the information word and codeword in terms of the delay operator D:

$$\mathbf{u}(D) = \mathbf{u}_0 + \mathbf{u}_1 D + \dots + \mathbf{u}_{L-1} D^{L-1}$$
(9)

$$\mathbf{v}(D) = \mathbf{v}_0 + \mathbf{v}_1 D + \dots + \mathbf{v}_{L-1} D^{L-1}$$
(10)

Each convolutional encoder can also be described by a generator matrix G(D) using the delay operator. For example, the encoder in Fig. 1 has generator matrix

 $G(D) = (1 + D + D^2 \ 1 + D^2)$. If the convolutional encoder has feedback, the entries of the generator matrix are rational functions. Using the delay operator, we can write the tailbiting encoding procedure in compact form as

$$\mathbf{v}(D) \equiv \mathbf{u}(D)G(D) \pmod{1+D^L}$$
(11)

where G(D) is a rational generator matrix.

Example 3. We repeat the calculations in Example 2 using (11). The generator matrix is $G(D) = (1 + D + D^2 1 + D^2)$, and the information sequence $u(D) = 1 + D^3$ gives the codeword

$$\mathbf{v}(D) \equiv (1+D^3)(1+D+D^2\ 1+D^2)$$

$$\equiv (D^2+D^3\ 1+D+D^2+D^3)$$

$$\equiv (01)+(01)D+(11)D^2+(11)D^3\ (\text{mod}\ 1+D^4)$$

(12)

Example 4. Consider tailbiting of length L = 4 using the (systematic) feedback encoder shown in Fig. 5. It is equivalent to the encoder in Fig. 1 and its generator matrix is

$$G(D) = \left(1\frac{1+D^2}{1+D+D^2}\right)$$
(13)

(14)

The information sequence $u(D) = 1 + D + D^3$ gives the codeword

$$\mathbf{v}(D) \equiv (1+D+D^3) \left(1 \ \frac{1+D^2}{1+D+D^2} \right)$$
$$\equiv (1+D+D^3)(1 \ (1+D^2)(1+D^2+D^3))$$
$$\equiv (1+D+D^3)(1 \ D+D^3) \equiv (1+D+D^3 \ D+D^3)$$
$$\equiv (10) + (11)D + (00)D^2 + (11)D^3 \ (\text{mod} \ 1+D^4)$$

where we have assumed that $(1 + D + D^2)^{-1} \equiv 1 + D^2 + D^3 \pmod{1 + D^4}$.

For certain encoders at certain tailbiting lengths the tailbiting technique will fail to work; we do not have a one-to-one mapping between the information sequences and the codewords.

Example 5. Assume that the feedback encoder in Fig. 5 is used for tailbiting. We start the encoder in state (0 1)



Figure 5. A rate $R = \frac{1}{2}$ (systematic) feedback convolutional encoder.

and feed it with zeros. The encoder passes the states $(1\ 0)$ and $(1\ 1)$ and after three steps the encoder will be back again in state $(0\ 1)$. The corresponding encoder output is 00 01 01. If the tailbiting length L is a multiple of 3, the repetition of this cycle L/3 times will be a tailbiting path corresponding to a nonzero codeword. Since the all-zero codeword always corresponds to the all-zero input, we will have more than one codeword corresponding to an all-zero input; that is, one information sequence corresponds to at least two codewords; hence, tailbiting fails when L is a multiple of 3.

The generator matrix G(D) used to generate a tailbiting code of length L can be decomposed in its invariant factor decomposition $G(D) = A(D)\Gamma(D)B(D)$ (see, for example, [6]), where A(D) and B(D) are $b \times b$ and $c \times c$ binary polynomial matrices with unit determinants, respectively, and where $\Gamma(D)$ is the $b \times c$ matrix

$$\Gamma(D) = \begin{pmatrix} \frac{\gamma_1(D)}{q(D)} & & & \\ & \ddots & & \\ & & \frac{\gamma_b(D)}{q(D)} & 0 & \dots & 0 \end{pmatrix}$$
(15)

where q(D) and $\gamma_i(D)$, i = 1, 2, ..., b, are binary polynomials. It was shown [4] that if and only if both q(D) and $\gamma_b(D)$ are relatively prime to $1 + D^L$, then (11) describes a one-to-one mapping between $\mathbf{u}(D)$ and $\mathbf{v}(D)$ for all $\mathbf{u}(D)$.

Example 6. Consider the feedback encoder in Fig. 5 with generator matrix G(D) given in (13). Using the invariant factor decomposition, we can write

$$G(D) = (1) \left(\frac{1}{1+D+D^2} \ 0\right) \left(\begin{array}{cc} 1+D+D^2 & 1+D^2 \\ 1+D & D \end{array}\right)$$
(16)

and, hence, $q(D) = 1 + D + D^2$ and $\gamma_b(D) = 1$. Since q(D) divides $1 + D^{3k}$, where k = 1, 2, ..., it follows in agreement with Example 5 that the tailbiting technique fails for L = 3k.

4. DECODING TAILBITING CODES

Assume that the information block \mathbf{u} is encoded as the codeword \mathbf{v} and then transmitted over a noisy channel. At the receiver side, the decoder is provided with a noise-corrupted version of \mathbf{v} that we denote by \mathbf{r} . The task of the decoder is to make an estimate $\hat{\mathbf{u}}$ of the information block \mathbf{u} based on the received \mathbf{r} . A decoder that chooses $\hat{\mathbf{u}}$ in such a manner that

$$\hat{\mathbf{u}} = \arg\max \mathbf{P}(\mathbf{r} \mid \mathbf{u} = \mathbf{x}) \tag{17}$$

where the maximum is taken over all possible information blocks, is called a (sequence) maximum-likelihood (ML) decoder. If all possible information blocks are equally likely to occur, then an ML decoder minimizes the probability that $\hat{\mathbf{u}} \neq \mathbf{u}$; thus, an ML decoder minimizes the decoding block error probability. A drawback of the ML decoder is

that it does not provide any information on how reliable the estimation $\hat{\bm{u}}$ is. A decoder that computes

$$P(u_t^{(i)} = x \mid \mathbf{r}), x \in \{0, 1\}$$

for all *t* and *i*, is called a (symbol-by-symbol) *a posteriori* probability (APP) decoder. The estimated information bit $\hat{u}_t^{(i)}$ is then chosen to be the binary digit *x* maximizing the probability $P(u_t^{(i)} = x | \mathbf{r})$; the probability $P(u_t^{(i)} = x | \mathbf{r})$ also provides valuable information on the reliability of the estimation $\hat{u}_t^{(i)}$, which is crucial in many modern coding schemes, including Turbo codes.

A code described via a trellis with only one possible starting state (as the example in Fig. 3b) may be decoded by trellis-based algorithms that can easily make use of the received soft channel output. Examples of such algorithms are the Viterbi decoder [7], which is an ML decoder, and the two-way [(BCJR) (Bahl-Cocke-Jelinek-Raviv)] decoder [8], which is an APP decoder. The main problem when decoding tailbiting codes is that the decoder does not know the encoder starting-ending state (since the starting-ending state depends on the information sequence to be transmitted), and, hence, the Viterbi decoder and two-way decoder cannot be used in their original forms. These algorithms, however, can be applied to tailbiting codes described via their tailbiting trellises with n_s possible encoder starting-ending states as follows. Let us divide the set of tailbiting codewords into n_s disjoint subsets or subcodes $C_{\mathbf{s}}^{tb}$, where $\mathbf{s} \in S$ and $C_{\mathbf{s}}^{tb}$ is the set of codewords that correspond to trellis paths starting and ending in the state ${\bf s}.$ Hence, each subcode is characterized by its encoder starting-ending state and corresponds to a trellis with one starting and one ending state. Each of these subcodes may be decoded by the Viterbi algorithm producing n_s candidate estimates of **u**. An ML decoder for tailbiting codes then chooses the best one, namely, the one with highest probability, among these n_s candidates as its output. Using a similar approach, we can also implement an a posteriori probability (APP) decoder [6].

The complexities of the described ML and APP decoders for tailbiting codes are of the order of n_s^2 . For a rate $R_{\rm tb} = b/c$ tailbiting code with memory m the number of possible encoder states n_s can be as large as 2^{mb} , and hence the complexity of the ML and APP decoders are then of the order of 2^{2mb} . Because of this high computational complexity, suboptimal decoding algorithms, that despite their lower complexities achieve performances close to those of the ML and APP decoders, have been widely explored.

Many suboptimal decoding methods make use of the circular trellis representations of tailbiting codes. The basic idea to circumvent the problem of the unknown encoder starting-ending state is to go around and around the circular trellis while decoding until a decoding decision is reached (see Fig. 6). The longer the decoder goes around the circular trellis, the less influence the unknown encoder starting state has on the performance of the decoder. There are many near-ML decoding algorithms using the Viterbi algorithm when circling around the tailbiting trellis; a nonexhaustive list of such decoders



Figure 6. Illustration of a suboptimal decoding algorithm going around and around a circular trellis.

is given by Calderbank et al. [9]. An approximate APP decoding algorithm using the two-way (BCJR) algorithm when circling around the tailbiting trellis was introduced by Anderson and Hladik [10]. Experiments have shown that often very few decoding cycles suffice to achieve satisfactory results. The decoding performance may, however, be significantly affected by *pseudocodewords* corresponding to trellis paths of more than one cycle that do not pass through the same trellis state at any integer multiple of the cycle length other than the pseudocodeword length. The described suboptimal tailbiting decoders going around and around the circular trellis have computational complexities of the order of 2^{mb} , which is the square root of the decoding complexities of the optimal algorithms.

5. APPLICATIONS OF TAILBITING CODES

The class of tailbiting codes includes many powerful block codes, as, for example, the binary (24, 12, 8) Golay code [9]. Furthermore, it was shown [11] that many of the best tailbiting codes have minimum distances as large as the best linear codes. Since for tailbiting codes regular trellis structures are directly obtained from their generators, these codes are often used when communicating over softoutput channels.

Using tailbiting codes as component codes in concatenated coding schemes is often an attractive option. For example, concatenated codes with outer Reed-Solomon codes and inner tailbiting codes have been adopted in the IEEE Standard 802.16-2001 for local and metropolitan area networks and in the ETSI EN 301 958 standard for digital videobroadcasting (DVB). Since tailbiting codes are easily decoded by (approximate) APP decoders, tailbiting codes are attractive component codes in concatenated coding schemes using iterative decoders [5].

BIOGRAPHIES

Marc Handlery received the Diploma degree in electrical engineering from the Swiss Federal Institute of Technology (ETH) Zúrich, Switzerland, in 1999. He is currently working at the Department of Information Technology at Lund University, Lund, Sweden, where he also received the Ph.D. degree in 2002. His research interests include convolutional codes, concatenated codes, tailbiting codes, and their trellis representations.

Rolf Johannesson received the M.S. and Ph.D. degrees in 1970 and 1975, respectively, both from Lund University, Lund, Sweden. He was awarded the degree of Professor, honoris causa, from the Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, in 2000. Since 1976, he has been with Lund University, where he is now Professor of Information Theory. His scientific interests include information theory, errorcorrecting codes, and cryptography. In addition to papers and book chapters in the area of convolutional codes and cryptography, he has authored two textbooks on switching theory and digital design and one on information theory and coauthored *Fundamentals of Convolutional Coding*.

Per Ståhl received his M.S. degree in electrical engineering in 1997 and his Ph.D. degree in information theory in 2001 from Lund University, Sweden. His research interests include tailbiting trellises and structure problems in error correcting codes. Since January 2002 he has been working with data security at Ericsson Mobile Platforms AB in Lund.

BIBLIOGRAPHY

- G. Solomon and H. C. A. van Tilborg, A connection between block and convolutional codes, SIAM J. Appl. Math. 37: 358–369 (Oct. 1979).
- H. H. Ma and J. K. Wolf, On tail biting convolutional codes, IEEE Trans. Commun. COM-34: 104-111 (Feb. 1986).
- G. D. Forney, Jr., *Review of Random Tree Codes*, NASA Ames Research Center, Contract NAS2-3637, NASA CR 73176, Final Report, Appendix A, Dec. 1967.
- P. Ståhl, J. B. Anderson, and R. Johannesson, A note on tailbiting codes and their feedback encoders, *IEEE Trans. Inform. Theory* 529-534 (Feb. 2002).
- C. Weiss, C. Bettstetter, and S. Riedel, Code construction and decoding of parallel concatenated tail-biting codes, *IEEE Trans. Inform. Theory* **IT-47**: 366–386 (Jan. 2001).
- R. Johannesson and K. Sh. Zigangirov, Fundamentals of Convolutional Coding, IEEE Press, Piscataway, NJ, 1999.
- A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* IT-13: 260-269 (April 1967).
- L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284-287 (March 1974).
- A. R. Calderbank, G. D. Forney, Jr., and A. Vardy, Minimal tail-biting trellises: The Golay code and more, *IEEE Trans. Inform. Theory* IT-45: 1435–1455 (July 1999).
- J. B. Anderson and S. M. Hladik, Tailbiting MAP decoders, IEEE J. Select. Areas Commun. 16: 297–302 (Feb. 1998).
- I. E. Bocharova, R. Johannesson, B. D. Kudryashov, and P. Ståhl, Tailbiting codes: Bounds and search results, *IEEE Trans. Inform. Theory* 48: 137–148 (Jan. 2002).

TELEVISION AND FM BROADCASTING ANTENNAS

HARUO KAWAKAMI YASUSHI OJIRO Antenna Giken Corp. Laboratory Saitama City, Japan

1. INTRODUCTION

At the present time, the superturnstile antenna (of which the batwing antenna is a radiating element) [1], the supergain antenna (dipole antennas with a reflector plate), invented in the United States; and the Vierergruppe antenna, invented in Germany and sometimes called the *two-dipole antenna* in Japan, are widely used for very high-frequency television (VHF-TV) broadcasting around the world. The superturnstile antenna (2) is used in Japan, as well as in the United States.

Figure 1 shows the appearance of the original batwing antenna, when it was first made public by Masters [1]. The secret lies in its complex shape. It is called a *batwing antenna* in the United States and *Schmetterlings antenne* (butterfly antenna) in Germany, in view of this shape.

The characteristics of the batwing antenna were calculated using the moment method proposed by Harrington [3], who conducted experiments on this antenna. The model antenna was approximately about



Figure 1. Historical shape of the batwing radiator.

two-fifths the size of a full-scale batwing antenna, with its design center frequency at 500 MHz.

It has been reported that for thick cylindrical antennas, a substantial effect on the current distribution appears to be due to nonzero current on the flat end faces. The assumption of zero current at the flat end face is appropriate for a thin cylindrical antenna; however, in the case of a thick cylindrical antenna, this assumption is not valid.

Specifically, this article applies the moment method to a full-wave dipole antenna, with a reflector plate supported by a metal bar, such as those widely used for TV and frequency-modulated (FM) broadcasting [4]. In the present study, the analysis is made by including the flat end-face currents. As a result, it is found that the calculated and measured values agree well, and satisfactory wideband characteristics are obtained.

Next, the twin-loop antennas, most widely used for ultra-high-frequency (UHF)- TV broadcasting, are considered. Previous researchers analyzed them by assuming a sinusoidal current distribution. Others adopted the higher-order expansions (Fourier series) of the current distribution; however, their analyses did not sufficiently explain the wideband characteristics of this antenna.

This article applies the moment method to a twin-loop antenna with a reflector plate or a wire-screen-type reflector plate. As for the input impedance, 2L-type twin-loop antennas have reactance near zero [in the case where $l_1 =$ $0.15\lambda_0$, i.e., where the voltage standing-wave ratio (VSWR) is nearly equal to unity]. Also, satisfactory wideband characteristics are obtained. The agreement between the measurement and the theory is quite good. Thus it may become possible in the future to improve practical antenna characteristics, based on the results obtained.

The digital terrestrial broadcasting station will use multiple channels in common. Therefore, antennas of the transmitting station will have the required properties of broad bandwidth and high gain. We have investigated the two-element modified batwing antenna with the reflector (UMBA) for UHF digital terrestrial broadcasting. It is calculated with the balanced feedshape. As a result, the broadband input impedance of 100 is obtained. However, in practical use, this antenna requires the balun and the impedance transformer as another circuit. So it has a complex feed system and high cost. In this section, we propose the unbalance-fed modified batwing antenna (UMBA), which does not use a balun and an impedance matching circuit. The same broad bandwidth and high gain compared with the previous feedshape are obtained. Next, the parallel coupling UMBA (PCUMBA) to obtain high gain is investigated. The gain of about 14 dBi is obtained for coupling four elements.

2. MOMENT METHOD

2.1. Thin Wire

This section discusses the moment method proposed by Harrington [3] and deals with the Garlerkin method where antenna current is developed using a triangle function and the same weight function as the current development function is used. The Garlerkin method can save calculation time because the coefficient matrix

2518 TELEVISION AND FM BROADCASTING ANTENNAS

is symmetric, and the triangle function is widely used because it presents a better performance in calculation time and accuracy and in other parameters for an antenna element without sudden change in antenna current.

Scattering electric field E^s by antenna current and charge is given by

$$E^s = -j\omega A - \nabla\phi \tag{1}$$

Vector potential A and scalar potential ϕ are given by

$$A = \frac{\mu}{4\pi} \iint_{S} J \frac{e^{-jkr_0}}{r_0} dS$$
 (2a)

$$\phi = \frac{1}{4\pi\varepsilon} \iint_{S} \sigma \frac{e^{-jkr_0}}{r_0} \, dS \tag{2b}$$

The following relation exists between charge density σ and current *J*:

$$\sigma = \frac{-1}{j\omega} \nabla \cdot J \tag{2c}$$

Now, assuming that the surface of each conductor is a perfect conductor and letting E^i be the incoming electric field, the following equation must hold true:

$$n \times E^S = -n \times E^i \tag{3}$$

The following simultaneous equation holds true from the boundary condition of the antenna surface of this antenna system:

$$\sum_{i=1}^{N} I_i \langle W_j, LF_i \rangle = \langle W_j, E_{\tan}^i \rangle,$$
$$\times (i = 1, \dots, N, \quad j = 1, \dots, M)$$
(4)

where L is the operator for integration and differentiation. The current is given, from Eq. (4), by

$$[I_i] = [\langle W_j, LF_i \rangle]^{-1} [\langle W_j, E_{\tan}^i \rangle]$$
(5)

and the matrix representation of (5) gives:

$$[I] = [Z]^{-1}[V] \tag{6}$$

Now, let the current at point t on each element be represented by

$$I(t) = \hat{t} \sum_{i=1}^{N} I_i T_i(t)$$
(7)

where \hat{t} is the unit vector in the direction of the antenna axis, and coefficient I_i is the complex coefficient determined by boundary condition. Letting $T_i(t)$ be the triangle development function, $T_i(t)$ is given by

$$T_{i}(t) = \begin{cases} 1 - \frac{|t - t_{i}|}{\Delta l_{i}} & t_{i-1} \langle t \langle t_{i+1} \rangle \end{cases}$$
(8)

where $\Delta l_i = t_i - t_{i-1}$ and $\Delta l_i = t_{i+1} - t_i$.

The impedance matrix Z in (6) is given by,

$$Z_{ji} = \int_{\text{axis}} dl \int_C dl' \left[j \omega \mu W_{jm} F_{\text{in}} + \frac{1}{j \omega \varepsilon} \frac{dW_{jm}}{dl} \cdot \frac{dF_{\text{in}}}{dl'} \right] \frac{e^{-jkr_0}}{4\pi r_0}$$
(9)

where C represents an antenna surface l' parallel to the antenna axis l. F_{in} and W_{jm} are divided into four in the triangle development function shown in Fig. 2, where the triangle is configured so that the value is one at the center and the divided antenna elements are obtained approximately by the four pulse functions. The expansion equation of (5) is used for the Green function.

2.2. Thick Wire

As shown in Fig. 3, a monopole antenna excited by a coaxial cable line consists of a perfectly conducting body of revolution being coaxial with the z axis. Conventionally, the integral equation is derived on the presumption that the tangential component of the scattered field cancels the corresponding impressed field component on the conductor surface.

Alternately, according to the boundary condition proposed by Dr. P. C. Waterman [6], an integral equation can





Figure 2. Approximation to the expansion and weighting function: (a) triangle function; (b) expansion and weighting function; (c) approximation to the expansion function F_{in} .



Figure 3. Coaxial cable line feeding a monopole through a ground plane and mathematical model of the antenna.

be derived by utilizing the field behavior within the conductor. Also, the conventional integral equation has a singular point when the source and the observation point are the same. On the other hand, the integral equation, after Waterman, is well behaved, and so it is more convenient for numerical calculation. When applying the extended boundary condition to an antenna having an axial symmetry as shown in Fig. 3, the axial component of the electric field is required to vanish along the axis of the conductor.

More explicitly, on the axis inside the conductor, the axial component of the total field is the sum of the scattered field (the field from current on the conductor) and the impressed field (the field from excitation). The corresponding integral equation is written as

$$\frac{j\eta}{4\pi k} \int_{-h}^{h} I_{z}(z') \left(k^{2} + \frac{\partial^{2}}{\partial z^{2}}\right) G(z, z') dz = E_{z}^{\text{inc}}$$
$$I_{z}(\pm h) = 0 \qquad (10)$$

with

$$\begin{split} G(z,z') &= \frac{e^{-jk\sqrt{(z-z')^2+a^2}}}{\sqrt{(z-z')^2+a^2}}, \quad k = \frac{2\pi}{\lambda}, \\ \eta &= 120\pi, \qquad I_z = 2\pi a J_z \end{split}$$

This integral equation reduces to the well-known equation (11) when the current is assumed to be zero on the antenna end faces:

$$\frac{j\eta}{4\pi k} \left\{ k^2 \int_S J_z(z') G(z,z') \, ds' + \frac{\partial}{\partial z} \int_S \nabla' \cdot J G(z,z') \, ds' \right\} = E_S^{\text{inc}}$$
(11)

Taking the current at the end faces into account, (11) becomes

$$\frac{j\eta}{4\pi k} \left\{ k^2 \int_{-h}^{h} I_z(z') G(z, z') dz' + \frac{\partial}{\partial z} \int_{-h}^{h} \frac{\partial I_z(z')}{\partial z'} G(z, z') dz' + \frac{\partial}{\partial z} \int_{S'} \frac{1}{\rho'} \frac{\partial J_\rho(\rho')}{\partial \rho'} G(z, z') ds' \right\} = E_S^{\text{inc}},$$

$$ds' = \rho' d\phi' d\rho' \tag{12}$$

where, on the end surface S', a in G(z, z') becomes ρ .

Dr. C. D. Taylor and Dr. D. R. Wilton [7] analyzed the current distribution on the flat end by a quasi-statictype approximation method; the resulting theoretical and experimental values were in good agreement. This analysis makes the same assumption that, as shown in Fig. 4, the current flowing axially to the center of the end surface without modification.

In applying the moment method, sinusoidal functions were used as expansion and weight functions. Accordingly, the Galerkin method was used to generate the integral equation.

Notice (Fig. 4) that the expansion and weight functions have been changed only on the dipole end faces. This is done so that the impedance matrix becomes symmetric for the end current and for the current flowing on the antenna surface (excluding the antenna end face).





Weighting function

Figure 4. Current and charge for flat end faces, expansion, and weighting function.

Expanding the unknown current in terms of sine functions, we have

$$I_z(z') = \sum_{n=1}^{N} I_n F_n$$
 (13)

If the expansion functions are overlapped

$$F_{n} = \begin{cases} \frac{\sin k(\Delta - |z' - z_{n}|)}{\sin k\Delta} & z' \in (z_{n-1}, z_{n+1}) \\ 0 & z' \notin (z_{n-1}, z_{n+1}) \end{cases}$$
$$\Delta = |z_{n+1} - z_{n}| \tag{14}$$

where z' indicates an axial coordinate taken along the conductor surface. Equations (13) and (14) are substituted into (12) to obtain

$$\frac{j\eta}{4\pi k \sin k\Delta} \sum_{n=1}^{N} I_n \left[\int_{Z_{n-1}}^{Z_{n+1}} G(z, z') \left(k^2 + \frac{d}{dz'^2} \right) F_n dz' \right. \\ \left. + k \{ G(z, z_{n+1}) + G(z, z_{n-1}) - 2\cos k\Delta G(z, z_n) \} \right] \\ \left. + \frac{j\eta}{4\pi k} \frac{\partial}{\partial z} \int_{S'} \frac{1}{\rho'} \frac{\partial J_\rho(\rho')}{\partial \rho'} G(z, z') ds' = E_z^{\text{inc}}$$
(15)

The first term in the integral is zero because F_n consists of sine functions. Assuming that

$$E_{z}^{\text{end}} = \frac{j\eta}{4\pi k} \frac{\partial}{\partial z} \int_{S'} \frac{1}{\rho'} \frac{\partial J_{\rho}(\rho')}{\partial \rho'} G(z, z') \, ds' \tag{16}$$

2520 TELEVISION AND FM BROADCASTING ANTENNAS

then (15) becomes

$$\frac{j\eta}{4\pi \sin k\Delta} \sum_{n=1}^{N} + I_n \{ G(z, z_{n+1}) + G(z, z_{n-1}) - 2\cos k\Delta G(z, z_n) \} + E_z^{\text{end}} = E_z^{\text{inc}}$$
(17)

Applying a weight function of the same form as (13) and taking the inner product of both sides of (17) and the weight function, we see that the equation becomes

$$\begin{aligned} \frac{j\eta}{4\pi \sin k\Delta} \sum_{n=1}^{N} I_n \langle G(z, z_{n+1}) + G(z, z_{n-1}) \\ &- 2\cos k\Delta G(z, z_n), W_m \rangle \\ &+ \langle E_z^{\text{end}}, W_m \rangle = \langle E_z^{\text{inc}}, W_m \rangle \quad (m = 1, 2, \dots N) \ (18) \end{aligned}$$

where

$$W_{m} = \begin{cases} \frac{\sin k(\Delta - |z - z_{m}|)}{\sin k\Delta} & z \in (z_{m-1}, z_{m+1}) \\ 0 & z \notin (z_{m-1}, z_{m+1}) \end{cases}$$
$$\Delta = |z_{m+1} - z_{m}| \tag{19}$$

Here z indicates a coordinate taken on the axis. The above results can be expressed in matrix form as

$$[Z][I] = [V] \tag{20}$$

The impressed field E_z^{inc} of the [V] matrix is considered to be excited by a frill of magnetic current (8) across the aperture of the coaxial cable line feeding a monopole as shown in Fig. 3. In other words, assuming the field on the aperture of a coaxial cable line at z = 0 to be identical to that of a transverse electromagnetic (TEM) mode on the coaxial transmission line, the equivalent frill of magnetic current can be determined. Also, by considering the image of an excitation voltage as V_0 , and the inner diameter and outer diameter of the coaxial cable line by a and b, respectively, it follows that

$$E_{z}^{\rm inc} = \frac{V_0}{2l_n(b/a)} \left\{ \frac{e^{-jk\sqrt{z^2 + a^2}}}{\sqrt{z^2 + a^2}} - \frac{e^{-jk\sqrt{z^2 + b^2}}}{\sqrt{z^2 + b^2}} \right\}$$
(21)

2.3. Charge Density and Current on End Faces

Now, we proceed with the study of the current over the antenna end surface. We presume that the current is zero at the center to the edge (in the case of a thin cylindrical antenna, the end current is assumed to be zero as shown by the dotted line in Fig. 3). If we express the total charge on the end surface by Q and its radius by a, then the charge density, σ , and the current density, J_{ρ} , are given by

$$\sigma = \frac{Q}{\pi a^2}, \quad J_\rho = \frac{-j\omega Q}{2\pi a^2}\rho \tag{22}$$



Figure 5. Input admittance of thick cylindrical antenna.





Figure 6. Construction of the model batwing antenna and its coordinate system.

Also, by the current continuity condition on the edge, we have $\mathbf{r} \in \mathcal{O}$

$$Q = \pm \frac{I_z(z')}{j\omega} \tag{23}$$

Furthermore, the axial component of the field strength on the z axis, produced by sources on the end surfaces, is given by

$$E_{z} = \pm \frac{j\eta I_{z}(z')}{2\pi k a^{2}} (z - z') \left\{ \frac{e^{-jk\sqrt{(z - z')^{2} + a^{2}}}}{\sqrt{(z - z')^{2} + a^{2}}} - \frac{e^{-jk|z - z'|}}{|z - z'|} \right\}$$
(24)

We impose the boundary condition that the total axial field is zero in the range of $-h\langle z \langle h, \text{ not including } z = \pm h$. Now, if we select the end face weight functions to be the same as the expansion functions, it follows that the same weight function will also be applied to the end faces $z = \pm h$, as shown in Fig. 4.

Referring to the difference between the analysis of Taylor and Wilton and that presented here, the former uses the point-matching method and is applied only to the body of revolution. On the other hand, the latter employs the Galerkin method and is applied to antennas that are asymmetric and that also contain discontinuities in the conductors. By using the moment method and taking the end surfaces into consideration, a relatively thick antenna can be treated.

For example, the calculated input admittance for $a = 0.0423\lambda$ and b/a = 1.187 is given in Fig. 5, as a function of h. For comparison, the values measured by Holly [9] are also shown. In the calculation, the number of subsections, N, was chosen as 50–60 per wavelength. The theoretical values agree well with the measured values. Therefore, it is concluded that the analytic technique is adequate for thick cylindrical antennas.

3. THE BATWING ANTENNA ELEMENT

The antenna is installed around a support mast, as shown in Fig. 6, and fed from points f and f', through a jumper from a branch cable with a characteristic impedance of 72 Ω . The conducting support mast is idealized by an infinite, thin mast. The batwing antenna element is divided into 397 segments for the original type, with triangular functions as the weighting and expansion functions, and the analysis of the batwing antenna elements is carried out using the Galerkin method. The batwing antenna is fed with unit voltage. The currents flowing in each antenna conductor are calculated over a frequency range of 300–700 MHz.

Figure 7a,b illustrates these current distributions $I_i(i = 1 \sim 12)$ on the conductors at frequencies of 300, 500, and 700 MHz. Since the distribution of currents along each conductor is calculated, this allows calculation of the radiation characteristics. Figure 8 illustrates the amplitude and phase characteristics of radiation patterns in the horizontal and vertical planes. It is seen from this figure that the theoretical values agree well with the measurements.

Figure 9 illustrates the theoretical and measured input impedance of a batwing antenna mounted on an aluminum

plate, $3m \times 3m$. Both curves coincide closely with each other, with the input impedance having a value close to 72 Ω , which is the proper match to the characteristic impedance of the branch cable. Vernier impedance matching is carried out in practice by connecting a metal jumper between the end of the branch cable and the feed point of the antenna element or the support mast. The feed strap's length, width, or form is varied to derive VSWR values below 1.10. In the case of the distance between the support mast and antenna elements being varied, Fig. 10 shows that the real part of the input impedance changes as the distance *l* between the support mast and antenna elements is varied.

Figure 11 shows that the real part of the input impedance is not largely affected by the angle of the jumper and that the reactance is likely to shift as a whole to the left side of the Smith chart because of the capacitance.

Figures 11 and 12 show that matching can be obtained by means of adjusting appropriately the distance between the support mast and antenna elements l, and the angle of the jumper in order to reduce the input VSWR.

The power gain of the antenna at 500 MHz is calculated to be 3.3 dB. Figure 13 shows the gain of the antenna in the $\phi = 0^{\circ}$ direction as a function of the frequency, referenced to a half-wavelength dipole. Figure 14 illustrates three-dimensional amplitude characteristics of radiation patterns in the horizontal and vertical planes at each frequency. For example, Fig. 15 indicates the polarization pattern of the calculated performance characteristics.

4. MUTUAL RADIATION IMPEDANCE CHARACTERISTICS

Two units of the batwing antenna illustrated in Fig. 6 were stacked one above the other in the same plane to constitute a broadside array with a center-to-center separation distance d.

Measurements were made as follows:

- 1. Frequency was fixed at 500 MHz, with distance d as the parameter.
- 2. Distance was fixed at 60 cm (full wavelength), with frequency as the parameter.

In both 1 and 2, measurements of mutual impedance were made under both open- and short-circuit conditions. Figure 16 shows a schematic of the equipment used in the measurement setup. The data obtained are shown in Fig. 17a,b. The theoretical values agree well with the measurements as shown in Fig. 17a,b. The mutual impedance was below a few ohms when the distance between elements exceeded 0.8 λ .

5. THEORETICAL ANALYSIS OF METAL-BAR-SUPPORTED WIDEBAND FULL-WAVE DIPOLE ANTENNAS WITH A REFLECTOR PLATE

Wideband full-wave dipole antennas with a reflector plate supported by metal bars were invented in Germany. The construction is shown in Fig. 18. A full-wave dipole antenna is located in front of a reflector, and supported



Figure 7. (a) Amplitude characteristics of current distribution for frequency range from 300, 500, 700 MHz of shaded areas; (b) current distribution at 500 MHz.



Figure 8. Amplitude and phase characteristics of radiation patterns at 300, 500, 700 MHz (w. support mast).

directly by a metal bar attached to a reflector. This antenna was also analyzed by the moment method described previously.

Because the supporting bar (see Fig. 18) is metallic, leakage currents may cause degradation of the radiation characteristics. To calculate these effects, the radial component of field E_{ρ} must be taken into consideration. In other words, E_{ρ} is needed for the calculation of $Z_{m,n}$, as defined by inner products of the expansion functions on the supporting bar and weighting functions on the antenna element or on the parallel conductors. We assume the supporting bar to be separated from the feed point by a distance l_1 . Also, the radius of the supporting bar is fixed at a fourth of the radius of the antenna element (i.e., at $\lambda_0/100$), and then is varied to be $0.2\lambda_0$, $0.25\lambda_0$, and $0.3\lambda_0$. Figures 19–22 indicate various calculated performance



Figure 9. Theoretical and measured values of the input impedance as function of frequency (mast is infinite thin, $\alpha = 0^{\circ}$).



Figure 10. Changing of distance between the support mast (infinite thin, $\alpha = 0^{\circ}$) antenna element.

characteristics. Note that the leakage current due to the supporting bar is minimized for $f/f_0 = 0.7$, and that this current is substantial at other frequencies. The current distribution is shown only for the case of $l_1 = 0.25\lambda_0$.

6. CHARACTERISTICS OF 2L TWIN-LOOP ANTENNAS WITH INFINITE REFLECTOR

As shown in Fig. 23, a twin-loop antenna has the loops connected by a parallel line: The 2L, 4L, and 6L types are used, according to the number of loops. For actual use, a reactive load is provided by the trap at the top end, which also serves as the antenna support. The dimensions

used for this article are as follows: center frequency $f_0 = 750$ MHz (wavelength $\lambda_0 = 40$ cm), length of the parallel line part $2l_1 = \lambda_0/2$ ($l_1 = 10$ cm), $l_2 = \lambda_0/2(l_2 = 20$ cm), interval of the parallel line part, $d = \lambda_0/20$ (d = 2 cm), loop radius $b = \lambda_0/2\pi$ (b = 6.366 cm), distance from the reflector to the antenna $l_3 = \lambda_0/4(l_3 = 10$ cm), conductor diameter $\phi = 10$ mm, and top end trap $l_t = 0$ to $\lambda_0/4$, changed in intervals of $\lambda_0/16$. 2L twin-loop antennas were arranged in front of an *infinite* reflector, and calculations were executed in regard to the frequency characteristics of the trap length.

The radiation pattern of the 2L-type antenna is shown in Fig. 24. Up to $l_t = \lambda_0/8$, the main beam gradually



Figure 11. Changing of the shape of the jumper ($\ell = 3.10$ cm).

becomes sharper with increasing frequency, and it can be seen that the sidelobes increase. When l_t increases in this way to $\lambda_0/8$ and $\lambda_0/4$, the directivity becomes disturbed.

Figure 24 shows $l_1 = 0.15\lambda_0$ and $l_1 = 0.25\lambda_0$ characteristics of the radiation pattern in a polar display. The antenna gain for both lengths ($l_1 = 0.15\lambda_0$ and $l_1 = 0.25\lambda_0$) shows a small change of approximately 9.5 - 8.5 dB. The input impedance has a value very close to 50Ω , essentially the same as the characteristic impedance of the feed cable. As for the input impedance, the 2L twin-loop antenna has reactance nearest zero (for the case where $l_1 = 0.15\lambda_0$); that is, the VSWR is nearly equal to unity.

In the calculation above, the reflector was considered to be an infinite reflector, and the effect of the reflector on the antenna elements was treated by the image method. In the case of practical antennas, however, it is the usual practice to make the reflector finite, or consisting of several parallel conductors. Therefore, a calculation was executed for a reflector in which 21 linear conductors replaced the infinite reflector, as shown in Fig. 25.

The results are shown along with those for the infinitereflector case. On the basis of these results, it was concluded that no significant difference was observed in input impedance and gain between the infinite reflector case and the case where the reflector consisted of parallel conductors.

The wire-screen-type reflector plate had a height of $3\lambda_0(120 \text{ cm})$, a width of $\lambda_0(40 \text{ cm})$, and a wire interval of



Figure 12. Theoretical and measured values of the input impedance of batwing antenna with support $\ell = 3.10 \text{ cm } \alpha = 0^{\circ}$).

 $0.15\lambda_0$ (6 cm). The radiation pattern is shown in Fig. 26. With regard to the pattern in the horizontal plane, no difference was found in comparison with an infinite reflector, but a backlobe of approximately $-16~\mathrm{dB}$ exists to the rear of the reflector. The same figure also shows the phase characteristics. With regard to the pattern in the vertical plane, the phase shows a large change where the pattern shows a cut.

7. THE DIGITAL TERRESTRIAL BROADCASTING ANTENNAS

7.1. Unbalance-Fed Modified Batwing Antenna

The configuration of the unbalance-fed modified batwing antenna (UMBA) is shown in Fig. 27. The reflector is



Figure 13. Batwing antenna gain with $\lambda/2$ dipole.

assumed an infinite perfect conducting plane for the calculation.

Parallel lines are connected perpendicularly to the center of batwing radiators. One side of the parallel line is connected perpendicularly to the reflector; the other side is bent above the reflector and connected to the other parallel line. UMBA is excited by the unbalanced generator, which is assumed to be a coaxial cable, between the center of the line and the reflector. The center frequency is 500 MHz $(\lambda_0=600~mm)$ and the conductor diameter is $0.013\lambda_0.$ The center frame of the batwing radiator, that is the length of LW0 = $0.25\lambda_0$ and the spacing of d, acts as a quarter wavelength stab. When the total length of the radiating elements LW1, LW2, and LW3 in Fig. 27 is about $0.5\lambda_0$, a broad-bandwidth input impedance is obtained. Therefore, $LW1 = 0.167\lambda_0$, $LW2 = 0.063\lambda_0$, and $LW3 = 0.271\lambda_0$ are chosen. Two batwing radiators are set on the height of $H = 0.25\lambda_0$ above the reflector. The spacing of d and Hf are adjusted in the range from $0.016\lambda_0$ to $0.025\lambda_0$ to obtain the input impedance of 50 Ω . The element spacing of *D* is $0.63\lambda_0$. NEC-Win Pro is used in the calculation.

In Fig. 28 the measured and the calculated VSWR (50 Ω) for a two-element UMBA are shown. The bandwidth ratio of 30% (VSWR \leq 1.15) is obtained for both results. In Fig. 29 the power gain is shown. The gain of about 11 dBi and the broadband property are obtained, and measured results agree well with calculated results. In Fig. 30, the normalized radiation patterns in the vertical plane (Z-Y plane: $E\phi$) and the horizontal plane (Z-X plane: E_{θ}) at frequencies of 450, 500, and 550 MHz are shown. Good symmetric patterns on the horizontal plane are obtained. The half-power beamwidth for five frequencies is shown in Table 1.

7.2. Parallel Coupling UMBA

The configuration of the parallel coupling UMBA (PCUMBA) is shown in Fig. 31. Four batwing elements are coupled without a space. The outside elements (1 and



Figure 14. Three-dimensional amplitude characteristics of radiation patterns at 300, 500, 700 MHz.



Figure 15. Polarization pattern of the model batwing antenna at 500 MHz (without support mast).



Figure 16. Block diagram of experimental measurement setup.



Figure 17. (a) Mutual radiation impedance at 500 MHz; (b) mutual radiation impedance changing of distance $d = 1\lambda$.



Figure 18. Metal-bar-supported wideband full-wave dipole antennas with a reflector plate.

4) are cross-connected with the inside elements (2 and 3) by the strait wire. Then the outside element is fed by in-phase against the inside one. Because the input impedance of UMBA is about 50 Ω , an impedance transformer is required for PCUMBA matched with the coaxial cable. The length from the center of PCUMBA to the center of 2(3) is a quarter-wavelength. Therefore, by changing the height and the radius of the center feed line, the impedance is transformed as a quarter-wave transformer. In Fig. 32 the calculated and measured VSWR normalized by 50 Ω for PCUMBA are shown. For the calculation, the bandwidth ratio (less than 1.15) is 30%. For the measurement, after adjusting stabs, the bandwidth ratio (less than 1.05) is 30%. Thus a broadband antenna is realized with the simple feed structure. In Fig. 33 the calculated power gain is shown. The gain of about 14 dBi equivalent for the five-element dipole array is obtained at the center frequency. In Fig. 34 radiation patterns with the infinite ground plane are shown. The half-power beamwidth for five frequencies is shown in Table 2.

8. CONCLUSION

Previous researchers [1,2] have analyzed batwing antennas by approximating the current distribution as a sinusoidal distribution. Wideband characteristics are not obtained with a sinusoidal current distribution. In this article, various types of modified batwing antennas, as the central form of the superturnstile antenna system, were analyzed theoretically with the aid of the moment method. The results were compared to measurements in order to examine the performance of the antenna elements in detail.

It is also evident from this research that the shape of the jumper has a remarkable effect on the reactance of the input impedance, and that the distance between the support mast and the antenna element also markedly influences the resistance of this impedance. Thus, a satisfactory explanation is given with regard to the matching conditions. As a result, it was found that the calculated and the measured values agree well, and satisfactory wideband characteristics are obtained. Also, it was found that a broadside array spaced over a full wavelength is to be of the order of a few ohms for the designed frequency at 500 MHz, and there would be small mutual coupling effect. It has been made clear that a superturnstile antenna made by arranging these antennas in multiple bays has an arrangement with small mutual coupling effect.

Next, an analytic method and calculated results for the performance characteristics of a thick cylindrical antenna were presented. The analysis used the moment method and took the end face currents into account. The calculated results were compared with measured values, demonstrating the accuracy of the analytic method.

Using this method, a full-wave dipole antenna with a reflector supported by a metal bar was analyzed. The input impedance was measured for particular cases, thus obtaining the antenna dimensions for which the antenna input impedance permits broadband operation. In conclusion, wideband characteristics are not obtained with a one-bay antenna. The wideband characteristic is



Figure 19. Current distribution of metal-bar-supported full-wave dipole antennas (two-bay) with a wirescreen-type reflector plate.

obtained by means of the mutual impedance of the twobay arrangement. In the frequency region of $f/f_0 = 0.7$, the resistance of the input impedance is considered to be constant. In this case, the leakage current to the support bar is small. With regard to the radiation pattern, it was seen that a degradation of characteristics was caused by the metal-support bar.

It is noted that the present method should be similarly useful for analyzing antennas of other forms where the end face effect is not negligible.



Figure 20. Radiation pattern of metal-bar-supported full-wave dipole antennas (two-bay) with a wire-screen-type reflector plate.

Next, the twin-loop antennas were considered for use as wideband antennas. The analysis results for the 2L type showed that the change in the characteristics with a change in frequency becomes more severe with increasing trap length l_t , and the bandwidth becomes small, while a short trap length l_t shows a small change and a tendency for the bandwidth to become wide. For $l_t = 0$, a wide bandwidth for pattern and gain was obtained for the 2L type. The input impedance has a value very close to 50 Ω , essentially the same as the characteristic impedance of



Figure 21. Input impedance characteristics of metal-bar-supported full-wave dipole antennas (two-bay) with a wire-screen-type reflector plate.



Figure 22. Gain of metal-bar-supported full-wave dipole antennas (two-bay) with a wire-screen-type reflector plate.



Figure 23. Structure of 2L-type twin-loop antenna and its coordinate system for analysis.



Figure 24. Vertical radiation pattern of 2L-type twin-loop antenna.



Figure 25. Structure of 2L-type twin-loop antenna with a wire-screen-type reflector plate.



Wire screen-type reflector

Figure 26. Comparison between characteristics of 2L-type twin-loop antenna with a wire-screen-type reflector and with an infinite reflector.



Figure 27. Configuration of UMBA.



Figure 28. VSWR characteristics of UMBA (normalized impedance $Z_0 = 50\Omega$).



Figure 29. Power gain of UMBA.

the feed cable over a very wide frequency range. Thus, a satisfactory explanation was given with regard to the matching conditions.

The unbalance fed modified batwing antenna and the parallel coupling UMBA were also investigated. Broadband properties the same as the balance-fed type is obtained and the input impedance is matched ell with a coaxial cable of 50 Ω . The unbalanced current are decreased, and good symmetric patterns are obtained by the contribution of the quarter-wavelength stab on the batwing element. A power gain for the parallel coupling four-element PCUMBA of about 14 dBi is obtained. The unbalance-fed shape and the parallel coupling shape have the effect that the digital terrestrial broadcasting antenna is simpler, smaller, and of wider bandwidth than another transmitting antennas for UHF-TV broadcasting.

It has been reported here that a rigorous theoretical analysis has been achieved about 45 years after the invention of these VHF-UHF antennas.

BIOGRAPHIES

Haruo Kawakami received the B.E. degree from the Department of Electrical Engineering, Meiji University, Tokyo, Japan, in 1962, and the Ph.D. degree from Tohoku University, Sendai, Japan, in 1983.

In 1962, he joined YAGI Antenna Co., Ltd. In 1964, he become a research associate of at the Department



Figure 30. Radiation pattern of UMBA: (a) horizontal plane $(E\theta)$; (b) vertical plane $(E\phi)$.

Table 1. Half-Power Bean	nwidth of	UMBA
--------------------------	-----------	------

	$400 \mathrm{MHz}$	$450 \mathrm{~MHz}$	$500 \mathrm{~MHz}$	$550 \mathrm{~MHz}$	600 MHz
Horizontal plane $(E\theta)$ Vertical plane $(E\phi)$	$71^\circ \\ 45^\circ$	${64^\circ \over 39^\circ}$	69° 40°	73° 32°	73° 32°

Table 2. H	alf-Power	Beamwidth	of PCUMBA
------------	-----------	-----------	-----------

	$400 \mathrm{MHz}$	$450 \mathrm{~MHz}$	$500 \mathrm{~MHz}$	$550 \mathrm{~MHz}$	600 MHz
Vertical plane $(E\phi)$	34°	30°	24°	22°	22°



Figure 31. Configuration of PCUMBA.

of Electric and Electronic Engineering, Sophia University, and was promoted to lecturer and associate professor there in 1985 and 1992, respectively. He is currently a director of Antenna Giken Co., Ltd. In 1989, he was a part-time lecturer at the Department of Electronics Engineering, Tokyo Metropolitan College of Aeronautical Engineering. In 1998, he was a visiting professor at Utsunomiya University.

Dr. Kawakami is the author of Antenna Theory and Application, New Electrical Circuit Practice, and New Alternate Current Circuit Practice, Mimatsu Data, Tokyo, 1991, Kogaku-Tosho, Tokyo, 1990 and 1992, respectively.

He has been engaged in research and development of mobile systems, automobile-borne antennas, and electromagnetic compatibility. Dr. Kawakami received a Best Defense Technology Paper Award in 1987.



Figure 32. VSWR characteristics of PCUMBA (normalized impedance $Z_0 = 50\Omega$).



Figure 33. Power gain of PCUMBA.

He is a member of the Applied Computational Electromagnetic Society, the Institute of Electronics, Information and Communication Engineers of Japan, and the Institute of Image Information and Television Engineers of Japan. Kawakami's name has been listed in Marquis "Who's who in the World."

Yasushi Ojiro received the B.E. degree from the College of Engineering Sciences, Tsukuba University, Ibaraki, Japan, in 1993, and the Ph.D. degree from Tsukuba University in 1998. In 1998, he jointed Antenna Giken Co., Ltd. He has been engaged in research and development of HF antenna, TV broadcasting antenna, and adaptive array. He is a member of the Institute of Electonics, Information and Communication Engineers of Japan.



600MHz Figure 34. Radiation pattern of PCUMBA [vertical plane (E_{ϕ})].

BIBLIOGRAPHY

TERNARY SEQUENCES

1. R. W. Masters, The super turnstile, *Broadcast News* **42**: (Jan. 1946).

550MHz

- 2. Y. Mushiake, ed., Antenna Engineering Handbook, The OHM-Sha, Ltd., Oct. 1980 (in Japanese).
- 3. R. F. Harrington, *Field Computation by Moment Method*, Macmillan, New York, 1968.
- W. Berndt, Kombinierte sendeantennen fur fernseh-und UKWrunfunk (teil II), *Telefunken-Zeitung*, *Jahrgang* 101: (Aug. 1953).
- 5. R. F. Harrington, Matrix method for field problems, *Proc. IEEE* 55(2): 136 (Feb. 1967).
- P. C. Waterman, Matrix formulation of electromagnetic scattering, *Proc. IEEE* 53(8): 785 (Aug. 1965).
- C. D. Taylar and D. R. Wilton, The extended boundary condition solution of the dipole antenna of revolution, *IEEE Trans. Antennas Propag.* AP-20(6): 772 (Nov. 1972).
- L. L. Tsai, Numerical solution for the near and far field of an annular ring of magnetic current, *IEEE Trans. Antennas Propag.* AP-20(5): 569 (Sept. 1972).
- 9. R. W. P. King, *Tables of Antenna Characteristics*, IFI/Plenum, New York, 1971.

TOR HELLESETH University of Bergen Bergen, Norway

1. INTRODUCTION

Sequences with good correlation properties have many applications in modern communication systems. Applications of sequences include signal synchronization, navigation, radar ranging, random-number generation, spread-spectrum communications, multipath resolution, cryptography, and signal identification in multiple-access communication systems.

In *code-division multiple-access* (CDMA) systems several users share a common communication channel by assigning a distinct signature sequence to each user, enabling the user to distinguish his/her signal from those of other users. It is important to select the distinct signature sequences in a way that minimizes the interference from the other users. It is also important to select the sequences such that synchronization is easily achieved.

Phase shift keying (PSK) is a commonly used modulation scheme. In this case the code symbols are

pth roots of unity. Even though binary sequences (p = 2) are used in most applications, one can sometimes arrive at better results by going to a larger alphabet. In this article we will consider constructions of ternary sequences (p = 3). Ternary sequence families can be constructed with better correlation properties than is possible to obtain by using comparable binary sequence families.

In many cases the best ternary sequences belong to a family of sequences that can be constructed for any p. In other cases the ternary sequences do not have a known nonternary generalization. We will therefore often consider sequences over an alphabet of size being a prime p, when the best ternary sequences can be obtained in this way by letting p = 3.

Many sequence families with good correlation properties that are used in many applications employ maximumlength sequences, in short denoted as m sequences, as one of the main ingredients. We will therefore give a detailed definition and presentation of the properties of these sequences. This requires some background material on finite fields (or Galois fields), which will be provided for the sake of completeness.

The m sequences are perhaps the most well known sequences because of their many applications in communication and cryptographic systems. They are deterministic and easy to generate in hardware using linear shift registers and still they possess many randomlike properties. One important property is that an m sequence has a twolevel autocorrelation function, a fact that is very useful for synchronization purposes.

From m sequences one can construct ternary sequences with even better autocorrelation properties that are not achievable using binary sequences. Further, we present some more recent constructions of sequences with twolevel autocorrelation obtained from ternary m-sequences.

The important aspects for applications are the correlation properties of a sequence or a family of sequences. We define the auto- and cross-correlations of sequences and discuss the important design parameters required by a family of sequences that are to be applied in CDMA systems.

In many practical applications the correlations that occur are rather more aperiodic than periodic in nature. The problem of designing families of sequences with good aperiodic correlation values is very hard in general. Therefore a common approach has been to design sequences with good periodic correlations and see if they satisfy the requirements needed by the applications at hand. The focus in this article will therefore be on periodic correlation.

The main part will be to construct large families of ternary sequences with good correlation properties. Since many of the best sequence families are based on the properties of the cross-correlation between two msequences, this will be studied in detail.

In order to compare the different constructions of sequence families, we will describe the best known bounds, due to Sidelnikov [1] and Welch [2], on the correlation values of the families. These bounds indicate that nonbinary sequence families can have better correlation parameters than the best binary families. In later sections we will construct several families of sequences with the best known correlation properties among ternary sequences. These constructions will include sequences by Trachtenberg [3] that represent generalizations of the important binary sequences due to Gold [4]. Some of the constructed sequence families due to Sidelnikov [5], Kumar and Moreno [6], and Helleseth and Sandberg [7] can be shown to be asymptotically optimal. Finally, we will also include tables containing the parameters of some of the best ternary sequence families.

2. CORRELATION OF SEQUENCES

The correlation between two sequences $\{u(t)\}\$ and $\{v(t)\}\$ of length n is the complex inner product of one of the sequences with a shifted version of the other. The correlation is *periodic* if the shift is a cyclic shift, *aperiodic* if the shift is noncyclic, and a *partial-period* correlation if the inner product involves only a partial segment of the two sequences. We will concentrate on the periodic correlation.

Definition 1. Let $\{u(t)\}$ and $\{v(t)\}$ be two complex-valued sequences of period *n*, not necessarily distinct. The periodic correlation of the sequences $\{u(t)\}$ and $\{v(t)\}$ at shift τ is defined by

$$\theta_{u,v}(\tau) = \sum_{t=0}^{n-1} u(t+\tau)v^*(t)$$

where the sum over t is computed modulo n and * denotes complex conjugation.

When the two sequences are the same, the correlation is called the *autocorrelation* of the sequence $\{u(t)\}$, whereas when they are distinct, it is common to refer to the *crosscorrelation*.

Example 1. The sequence of period n = 13,

$$u(t) = 0 \ 0 - 1 \ 0 - 1 - 1 - 1 + 1 + 1 \ 0 - 1 + 1 - 1$$

has autocorrelation $\theta_{u,u}(\tau) = 0$ for all $\tau \neq 0 \pmod{n}$.

Example 2. Let ω be a complex third root of unity: $\omega^3 = 1$. The sequence of period n = 26

$$\begin{aligned} \{u(t)\} &= 11\omega^2 1\omega^2 \omega^1 \omega^2 \omega^2 \omega^1 1 \omega^2 \omega^2 \omega^2 11 \omega^1 1 \omega^1 \omega^2 \omega^1 \omega^1 \omega^2 \\ & \times 1\omega^1 \omega^1 \omega^1 \end{aligned}$$

has autocorrelation $\theta_{u,u}(\tau) = -1$ for all $\tau \neq 0 \pmod{n}$.

Frequently, as $\{u(t)\}$ in example 2, the sequences $\{u(t)\}$ and $\{v(t)\}$ are of the form $u(t) = \omega^{a(t)}$ or $v(t) = \omega^{b(t)}$, where ω is a complex *p*th root of unity and where the sequences $\{a(t)\}$ and $\{b(t)\}$ take values in the set of integers mod *p*. In this case we will also sometimes use the notation $\theta_{a,b}$ instead of $\theta_{u,v}$.

For synchronization purposes one prefers sequences with low absolute values of the maximum out-of-phase autocorrelation; that is, $|\theta_{u,u}(\tau)|$ should be small for all values of $\tau \neq 0 \pmod{n}$. For most applications one needs families of sequences with good simultaneously auto- and cross-correlation properties.

Let \mathcal{F} be a family consisting of M sequences

$$\mathcal{F} = \{\{s_i(t)\}: i = 1, 2, \dots, M\}$$

where each sequence $\{s_i(t)\}$ has period *n*.

The cross-correlation between two sequences $\{s_i(t)\}$ and $\{s_j(t)\}$ at shift τ is denoted by $C_{i,j}(\tau)$. In CDMA applications it is desirable to have a family of sequences with certain properties. To facilitate synchronization, it is desirable that all the out-of-phase autocorrelation values $(i = j, \tau \neq 0)$ are small. To minimize the interference due to the other users in a multiple-access situation, the crosscorrelation values $(i \neq j)$ must also be kept small. For this reason the family of sequences should be designed to minimize

$$C_{\max} = \max\{|C_{i,j}|: 1 \le i, j \le M, \text{ and either } i \ne j \text{ or } \tau \ne 0\}$$

For practical applications one needs a family \mathcal{F} of sequences of period n, such that the number of users $M = |\mathcal{F}|$ is large and simultaneously C_{\max} is small.

3. GALOIS FIELDS

Many sequence families are most easily described using the theory of finite fields. This section gives a brief introduction to finite fields. In particular the finite field $GF(3^3)$ will be studied in order to provide detailed examples of good sequence families.

There exists finite fields, also known as *Galois fields*, with p^m elements for any prime p and any positive integer m. A Galois field with p^m elements is unique (up to an isomorphism) and is denoted $GF(p^m)$.

For a prime p, let $GF(p) = \{0, 1, ..., p-1\}$ denote the integers modulo p with the two operations addition and multiplication modulo p.

To construct a Galois field with p^m elements, select a polynomial f(x) of degree m, with coefficients in GF(p) that is irreducible over GF(p); thus, f(x) can not be written as a product of two polynomials, of degree ≥ 1 , with coefficients from GF(p) [irreducible polynomials of any degree m over GF(p) exist].

Let

$$GF(p^{m}) = \{a_{m-1}x^{m-1} + a_{m-2}x^{m-2} + \dots + a_{0}; a_{0}, \dots, a_{m-1} \in GF(p)\}$$

Then $GF(p^m)$ is a finite field when addition and multiplication of the elements (polynomials) are done modulo f(x) and modulo p. To simplify the notations, let α denote a zero of f(x), that is, $f(\alpha) = 0$. Such an α exists, it can formally be defined as the equivalence class of x modulo f(x).

Example 3. The Galois field $GF(3^3)$ can be constructed as follows. Let $f(x) = x^3 + 2x + 1$, which is easily seen to be an irreducible polynomial over GF(3). Then $\alpha^3 = \alpha + 2$ and

$$GF(3^3) = \{a_2\alpha^2 + a_1\alpha + a_0; a_0, a_1, a_2 \in GF(3)\}$$

Computing the powers of α , we obtain

$$\begin{aligned} \alpha^4 &= \alpha \cdot \alpha^3 = \alpha(\alpha+2) = \alpha^2 + 2\alpha \\ \alpha^5 &= \alpha \cdot \alpha^4 = \alpha(\alpha^2+2\alpha) = \alpha^3 + 2\alpha^2 = 2\alpha^2 + \alpha + 2 \\ \alpha^6 &= \alpha \cdot \alpha^5 = \alpha(2\alpha^2+\alpha+2) = 2\alpha^3 + \alpha^2 + 2\alpha = \alpha^2 + \alpha + 1 \end{aligned}$$

and, similarly, all higher powers of α can be expressed as a linear combination of α^2 , α , and 1. In particular, the calculations give $\alpha^{26} = 1$. In Table 1 we give all the powers of α as a linear combination of 1, α , and α^2 . The polynomial $a_2\alpha^2 + a_1\alpha + a_0$ is represented as $a_2a_1a_0$.

Hence, the elements $1, \alpha, \alpha^2, \ldots, \alpha^{25}$ are all the nonzero elements in GF(3³). In general, such an element α , which generates the nonzero elements of GF(p^m), is called a *primitive element* in GF(p^m). Every finite field has a primitive element. An irreducible polynomial g(x) of degree m with coefficients in GF(p) with a primitive element as a zero is called a *primitive polynomial*.

All elements in $GF(p^m)$ are roots of the equation $x^{p^m} - x = 0$. Let β be an element of $GF(p^m)$. It is important to study the polynomials of smallest degree with coefficients in $GF(p^m)$ that has β as a zero. This polynomial is called the *minimum polynomial* of β over GF(p). Note that since all the binomial coefficients $\binom{p}{i}$ are divisible by p, for 1 < i < p, it follows for any $a, b \in GF(p^m)$ that

$$(a+b)^p = a^p + b^p$$

Observe that if $m(x) = \sum_{i=0}^{k} m_i x^i$ has coefficients in GF(p)

and β as a zero, then

$$m(\beta^p) = \sum_{i=0}^{\kappa} m_i \beta^{pi} = \sum_{i=0}^{\kappa} m_i^p \beta^{pi} = \left(\sum_{i=0}^{\kappa} m_i \beta^i\right)^p$$
$$= (m(\beta))^p = 0$$

Hence, m(x) has β , β^p , ..., $\beta^{p^{\kappa-1}}$, as zeros where κ is the smallest integer such that $\beta^{p^{\kappa}} = \beta$. Conversely, the polynomial with exactly these zeros can be shown to be an irreducible polynomial.

Example 4. We will find the minimal polynomial of all the elements in GF(3³). Let α be a root of $x^3 + 2x + 1 = 0$,

	The Galois Field GF(he Galois Field GF(3 ⁸	3)
--	----------------------	-----------------------------------	----

i	$lpha^i$	i	α^i
0	001	13	002
1	010	14	020
2	100	15	200
3	012	16	021
4	120	17	210
5	212	18	121
6	111	19	222
7	122	20	211
8	202	21	101
9	011	22	022
10	110	23	220
11	112	24	221
12	102	25	201

that is, $\alpha^3 = \alpha + 2$. The minimal polynomials over GF(3) of α^i for $0 \le i \le 25$ are denoted $m_i(x)$. Observe by the argument above that $m_{pi}(x) = m_i(x)$, where the indices are taken modulo 26. It follows that

$$\begin{split} m_0(x) &= (x - \alpha^0) = x + 2, \\ m_1(x) &= (x - \alpha)(x - \alpha^3)(x - \alpha^9) = x^3 + 2x + 1, \\ m_2(x) &= (x - \alpha^2)(x - \alpha^6)(x - \alpha^{18}) = x^3 + x^2 + x + 2, \\ m_4(x) &= (x - \alpha^4)(x - \alpha^{12})(x - \alpha^{10}) = x^3 + x^2 + 2, \\ m_5(x) &= (x - \alpha^5)(x - \alpha^{15})(x - \alpha^{19}) = x^3 + 2x^2 + x + 1, \\ m_7(x) &= (x - \alpha^7)(x - \alpha^{21})(x - \alpha^{11}) = x^3 + x^2 + 2x + 1, \\ m_8(x) &= (x - \alpha^8)(x - \alpha^{24})(x - \alpha^{20}) = x^3 + 2x^2 + 2x + 2, \\ m_{13}(x) &= (x - \alpha^{13}) = x + 1, \\ m_{14}(x) &= (x - \alpha^{14})(x - \alpha^{16})(x - \alpha^{22}) = x^3 + 2x + 2, \\ m_{17}(x) &= (x - \alpha^{17})(x - \alpha^{25})(x - \alpha^{23}) = x^3 + 2x^2 + 1. \end{split}$$

This also leads to a factorization into irreducible polynomials of $x^{27} - x$:

$$\begin{aligned} x^{27} - x &= x \prod_{j=0}^{25} (x - \alpha^j) \\ &= x m_0(x) m_1(x) m_2(x) m_4(x) m_5(x) m_7(x) \\ &\times m_8(x) m_{13}(x) m_{14}(x) m_{17}(x) \end{aligned}$$

Since a polynomial $m_i(x)$ is primitive if it is irreducible and its zeros are primitive elements, it follows that the polynomial $m_i(x)$ is primitive whenever $gcd(i, p^m - 1) = 1$. In general the number of primitive polynomials of degree m over GF(p) is $\phi(p^m - 1)/m$, where ϕ is Euler's ϕ function, that is, $\phi(n)$ is the number of integers i such that $1 \le i < n$ and gcd(i, n) = 1. Thus, in our example, $m_1(x), m_5(x), m_7(x), m_{17}(x)$ are the $\phi(26)/3 = 4$ primitive polynomials of degree 3 over GF(3).

4. THE TRACE FUNCTION

In order to describe sequences generated by a linear recursion, it is convenient to introduce the *trace function* from $GF(p^m)$ to GF(p). The *trace* function from $GF(3^3)$ to GF(3) is illustrated in Example 5.

Example 5. The trace function from $GF(3^3)$ to GF(3) is defined by

$$Tr(x) = x + x^3 + x^9$$

Since $x^{27} = x$ and $(x + y)^3 = x^3 + y^3$ holds for all $x, y \in GF(3^3)$, it is easy to see that $Tr(x) \in GF(3)$ for all $x \in GF(3^3)$, because

$$(\operatorname{Tr}(x))^3 = (x + x^3 + x^9)^3 = x^3 + x^9 + x^{27}$$

= $x^3 + x^9 + x = \operatorname{Tr}(x)$

Further, it follows that

$$Tr(x + y) = Tr(x) + Tr(y), Tr(ax)$$
$$= aTr(x), \text{ and } Tr(x^3) = Tr(x)$$

for any $x, y \in GF(3^3)$ and $a \in GF(3)$.

It is straightforward to calculate the trace of the elements in $GF(3^3)$:

Tr(1) = 1 + 1 + 1 = 0 $Tr(\alpha) = \alpha + \alpha^{3} + \alpha^{9} = 0$ $Tr(\alpha^{2}) = \alpha^{2} + \alpha^{6} + \alpha^{18} = 2$ $Tr(\alpha^{4}) = \alpha^{4} + \alpha^{12} + \alpha^{10} = 2$ $Tr(\alpha^{5}) = \alpha^{5} + \alpha^{15} + \alpha^{19} = 1$ $Tr(\alpha^{7}) = \alpha^{7} + \alpha^{21} + \alpha^{11} = 2$ $Tr(\alpha^{8}) = \alpha^{8} + \alpha^{24} + \alpha^{20} = 1$ $Tr(\alpha^{13}) = \alpha^{13} + \alpha^{13} + \alpha^{13} = 0$ $Tr(\alpha^{14}) = \alpha^{14} + \alpha^{16} + \alpha^{22} = 0$ $Tr(\alpha^{17}) = \alpha^{17} + \alpha^{25} + \alpha^{23} = 1$

Since $\operatorname{Tr}(x) = \operatorname{Tr}(x^3)$ for all $x \in \operatorname{GF}(3^3)$, the trace of all the elements $\operatorname{Tr}(\alpha^t)$ for $t = 0, 1, \dots, 25$ are given by

00202122102220010121120111

The sequence $\{\text{Tr}(\alpha^t)\}$ is a ternary *m* sequence of period $n = 3^3 - 1$, to be studied further in the next section. Note also that $\text{Tr}(\alpha^{t+13}) = \alpha^{13}\text{Tr}(\alpha^t) = 2\text{Tr}(\alpha^t)$.

A more general description of the trace function is given below. It is well known that $\operatorname{GF}(q^k) \subset \operatorname{GF}(q^m)$ if and only if k divides m (denoted by $k \mid m$). Let q be a power of a prime p, and let $m = ke, k, e \geq 1$. Then the *trace function* Tr_k^m is a mapping from the finite field $\operatorname{GF}(q^m)$ to the subfield $\operatorname{GF}(q^k)$ given by

$$\operatorname{Tr}_{k}^{m}(x) = \sum_{i=0}^{e-1} x^{q^{ki}}$$

Lemma 1. The trace function satisfies the following:

- (a) $\operatorname{Tr}_{k}^{m}(ax + by) = a\operatorname{Tr}_{k}^{m}(x) + b\operatorname{Tr}_{k}^{m}(y)$, for all $a, b \in \operatorname{GF}(q^{k}), x, y \in \operatorname{GF}(q^{m})$.
- (b) $\operatorname{Tr}_{k}^{m}(x^{q^{k}}) = \operatorname{Tr}_{k}^{m}(x)$, for all $x \in \operatorname{GF}(q^{m})$.
- (c) Let l be an integer such that k|l|m. Then

$$\operatorname{Tr}_{k}^{m}(x) = \operatorname{Tr}_{k}^{l}(\operatorname{Tr}_{l}^{m}(x)), \text{ for all } x \in \operatorname{GF}(q^{m}).$$

(d) For any $b \in GF(q^k)$, it holds that

$$|\{x \in \operatorname{GF}(q^m) \mid \operatorname{Tr}_{k}^{m}(x) = b\}| = q^{m-k}$$

(e) Let $a \in GF(q^m)$. If $Tr_k^m(ax) = 0$ for all $x \in GF(q^m)$, then a = 0.

2540 TERNARY SEQUENCES

In particular, it is useful to observe that the trace mapping $\operatorname{Tr}_{k}^{m}(x)$ takes on all values in the subfield $\operatorname{GF}(q^{k})$ equally often when *x* runs through $\operatorname{GF}(q^{m})$. In the cases when the fields involved are clear from the context, we will normally omit the subscripts.

5. MAXIMAL-LENGTH SEQUENCES (m SEQUENCES)

Maximal-length linear feedback shift register sequences (or *m*-sequences) are important in many applications and are building blocks for many important sequence families. An *m*-sequence has period $n = q^m - 1$ and symbols from a Galois field GF(q). During a period of an *m* sequence, each *m*-tuple of *m* consecutive symbols, except for the all zero *m*-tuple, occurs exactly once. Any *m* sequence can be generated from a linear recursion using a primitive polynomial. The following is an example of a ternary *m* sequence of period $n = 3^3 - 1 = 26$ having symbols from GF(3).

Example 6. Let f(x) be the ternary primitive polynomial defined by $f(x) = x^3 + 2x + 1$. Define the sequence $\{s(t)\}$ by $s(t+3) + 2s(t+1) + s(t) = 0 \pmod{3}$ i.e., $s(t+3) = s(t+1) + 2s(t) \pmod{3}$. Therefore, starting with the initial state (s(0), s(1), s(2)) = (002) one generates the *m* sequence

00202122102220010121120111...

of period $n = 3^3 - 1 = 26$. Three consecutive positions (s(t), s(t+1), s(t+2)) take on all possible nonzero values in GF(3)³ during a period of the sequence. All other nonzero initial states will generate a cyclic shift of this sequence.

Using the trace function, we can define a sequence $\{s(t)\}$ with symbols in GF(3) such that $s(t) = \text{Tr}(\alpha^t)$, where α is a zero of $f(x) = x^3 + 2x + 1$. This *m* sequence obeys the recursion $s(t + 3) + 2s(t + 1) + s(t) = 0 \pmod{3}$, since

$$s(t+3) + 2s(t+1) + s(t) = \operatorname{Tr}(\alpha^{t+3}) + 2\operatorname{Tr}(\alpha^{t+1}) + \operatorname{Tr}(\alpha^{t})$$
$$= \operatorname{Tr}(\alpha^{t+3} + 2\alpha^{t+1} + \alpha^{t})$$
$$= \operatorname{Tr}(\alpha^{t}(\alpha^{3} + 2\alpha + 1))$$
$$= \operatorname{Tr}(0)$$
$$= 0$$

All cyclic shifts of $\{s(t)\}$ are described by $\{s(t + \tau)\} = {\text{Tr}(c\alpha^t)}$, where $c = \alpha^{\tau}$.

We next describe m sequences in general over the alphabet GF(q). A common method to generate sequences is via linear recursions. A linear recursion over GF(q) is given by

$$\sum_{i=0}^{m} f_i s(t+i) = 0, \text{ for all } t$$
 (1)

where $f_i \in GF(q)$ for i = 0, 1, ..., m. The sequence $\{s(t)\}$ is completely determined by the recursion above and the

initial values $s(0), s(1), \ldots, s(m-1)$. The *characteristic* polynomial of the recursion is defined to be

$$f(x) = \sum_{i=0}^{m} f_i x^i$$

The maximum possible period of a sequence generated by a recursion of degree m is $q^m - 1$. This follows since m consecutive symbols uniquely determine the sequence and there are only $q^m - 1$ possible nonzero possibilities for m successive symbols. The maximum period $n = q^m - 1$ is obtained in the case when f(x) is a primitive polynomial.

Some important properties of m sequences are listed next. These properties are easily verified by the sequence in Example 6, but hold for all m sequences in general.

Lemma 2. Let $\{s(t)\}$ be an *m* sequence of period $q^m - 1$ over GF(q).

- (a) (Balance property) All nonzero elements occur equally often q^{m-1} times and the zero element occurs $q^{m-1} 1$ times during a period of the *m* sequence.
- (b) (Run property) As t varies over $0 \le t \le q^m 2$, the *m*-tuple

$$(s(t), s(t+1), \ldots, s(t+m-1))$$

runs through all the elements in $GF(q)^m$ exactly once, with the exception of the all-zero *m*-tuple, which does not occur.

(c) (Shift and add property) For any τ , $0 < \tau \le q^m - 2$, there exists a δ for which

$$s(t) - s(t + \tau) = s(t + \delta)$$
, for all t

(d) (Constancy on cyclotomic cosets) There exists a cyclic shift τ of $\{s(t)\}$, such that $s(p^i t + \tau) = s(t + \tau)$ for all t.

The *m* sequences generated by the different primitive polynomials $m_i(x)$ are described in Table 2. The *m* sequences generated by $m_i(x)$ are all cyclic shifts of the sequence $\{Tr(\alpha^{it})\}$, which is the one listed in Table 2. The sequence $\{s(dt)\}$ is said to be a decimation (by *d*) of the sequence $\{s(t)\}$, where indices are computed modulo the period *n* of $\{s(t)\}$. Further, from the trace representation, one observes that an *m* sequence generated by $m_i(x)$ is obtained by decimating an *m*-sequence generated by $m_1(x)$ by *i*.

Table 2. *m* Sequences of Period $n = 3^3 - 1 = 26$

i	$m_i(x)$	m Sequence
1	$x^3 + 2x + 1$	00202122102220010121120111
5	$x^3 + 2x^2 + x + 1$	01211120011010212221002202
$\overline{7}$	$x^3 + x^2 + 2x + 1$	02022001222120101100211121
17	$x^3 + 2x^2 + 1$	01110211210100222012212020

Definition 2. We define two sequences $\{s_1(t)\}$ and $\{s_2(t)\}$ to be cyclically equivalent if there exists an integer τ such that

$$s_1(t+\tau) = s_2(t)$$

for all *t*, otherwise they are said to be *cyclically distinct*.

The sequences generated by different $m_i(x)$ can be shown to be cyclically distinct. Therefore, there are $\phi(3^3 - 1)/3 = 4$ cyclically distinct *m* sequences of period 26.

6. SEQUENCES WITH LOW AUTOCORRELATION

One attractive property of *m* sequences of period $n = p^m - 1$, *p* a prime, is their *two-level autocorrelation* property.

Theorem 1. The autocorrelation function for an *m* sequence $\{s(t)\}$ of period $n = p^m - 1$, where *p* is a prime, is given by

$$\theta_{s,s}(\tau) = \begin{cases} -1 & \text{if } \tau \neq 0 \; (\text{mod } p^m - 1) \\ p^m - 1 & \text{if } \tau = 0 \; (\text{mod } p^m - 1) \end{cases}$$

Proof In the case $\tau = 0 \pmod{p^m - 1}$, the result follows directly from the definition. In the case $\tau \neq 0 \pmod{p^m - 1}$, define $u(t) = s(t + \tau) - s(t)$ and observe that $\{u(t)\}$ is a nonzero sequence. Further, $\{u(t)\}$ is an *m* sequence by the shift-add property. The balance property of an *m* sequence implies that

$$\begin{aligned} \theta_{s,s}(\tau) &= \sum_{i=0}^{p^m-2} \omega^{s(t+\tau)-s(t)} \\ &= \sum_{i=0}^{p^m-2} \omega^{u(t)} \\ &= (p^{m-1}-1)\omega^0 + p^{m-1} \sum_{i=1}^{p-1} \omega^i = -1 \end{aligned}$$

where ω is a complex primitive *p*th root of unity.

Other well-known sequences with two-level autocorrelation properties are the GMW sequences, due to Gordon, Welch, and Mills. These sequences can most easily be described in terms of the trace function.

Theorem 2. Let k, m be integers with $k \mid m, k \geq 1$. Let $r, 1 \leq r \leq q^k - 2$ satisfy $gcd(r, q^k - 1) = 1$. Let a be a nonzero element in $GF(q^m)$, and let $\{s(t)\}$ be the GMW sequence of period $q^m - 1$ defined by

$$s(t) = \operatorname{Tr}_{1}^{k} \{ [\operatorname{Tr}_{k}^{m}(a\alpha^{t})]^{r} \}$$

where α is a primitive element in $GF(q^m)$. Then

- (a) $\{s(t)\}$ is balanced.
- (b) If q is a prime, then

$$\theta_{s,s}(\tau) = \begin{cases} -1 & \text{if } \tau \neq 0 \; (\text{mod } q^m - 1) \\ q^m - 1 & \text{if } \tau = 0 \; (\text{mod } q^m - 1) \end{cases}$$

The low values of the out-of-phase autocorrelation of m sequences and GMW sequences make them attractive in many applications that require synchronization. The *linear span* of a sequence is the smallest degree of a characteristic polynomial that generates the sequence. One important advantage with GMW sequences is that they in general have a much larger linear span than do m sequences. This is important in some applications.

Definition 3. A sequence $\{s(t)\}$ of period *n* is said to have a "perfect" autocorrelation function if

$$\theta_{s,s}(\tau) = 0 \quad \text{for all } \tau \neq 0 \pmod{n}$$

Sequences with perfect autocorrelation functions do not always exist. For example, binary $\{-1, +1\}$ sequences with odd period cannot have this property since all the autocorrelation values necessarily have to be odd. For any period n > 4, no binary $\{-1, +1\}$ sequence is known to have perfect autocorrelation.

One family of *ternary* sequences $\{u(t)\}$ with perfect autocorrelation are the Ipatov sequences. These sequences are based on *m*-sequences over GF(q), *q* odd, of period $n = q^m - 1$, combined with the quadratic character of GF(q), which is a mapping from GF(q) to the set $\{0, -1, +1\}$.

A nonzero element x is said to be a square in GF(q)if $x = y^2$ has a solution $y \in GF(q)$. In order to define the ternary Ipatov sequences, first define the quadratic character χ :

$$\chi(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \text{ is a square in } \operatorname{GF}(q) \\ -1 & \text{if } x \text{ is a nonsquare in } \operatorname{GF}(q) \end{cases}$$

The quadratic character χ has the property that $\chi(xy) = \chi(x)\chi(y)$ for all $x, y \in GF(q)$. In particular, if α is a primitive element in GF(q), then the squares in GF(q) are exactly the even powers of α ; thus $\chi(\alpha^i) = (-1)^i$.

Combining the properties of m sequences and the quadratic character χ leads to the following ternary sequences where the out-of-phase autocorrelation is always zero. Note, however, that these sequences contain q^{m-1} zero elements.

Theorem 3. Let $\{s(t)\}$ be an *m* sequence over GF(q) of period $n = q^m - 1$ where $q = p^r$, *p* an odd prime, and *m* odd. Let f(x) be the characteristic polynomial of $\{s(t)\}$, and let α be a zero of f(x). The ternary sequence $\{u(t)\}$ with symbols from $\{0, -1, +1\}$ defined by

$$u(t) = (-1)^t \chi(s(t))$$

has period $N = (q^m - 1)/(q - 1)$ and perfect autocorrelation.

Proof The proof follows nicely from the properties of *m*-sequences and the quadratic character χ . First note that from the trace representation of an *m*-sequence it follows without loss of generality that

$$s(t+N) = \operatorname{Tr}(\alpha^{t+N}) = \alpha^N \operatorname{Tr}(\alpha^t)$$

where $\alpha^N \in GF(q)$. Hence

$$u(t+N) = (-1)^{t+N} \chi(s(t+N)) = (-1)^{t+N} \chi(\alpha^N) \chi(s(t))$$
$$= (-1)^t \chi(s(t)) = u(t)$$

which implies that $\{u(t)\}$ has period *N*.

To show that $\{u(t)\}$ has perfect autocorrelation, let $\tau \neq 0 \pmod{N}$ and use the property that $\{u(t)\}$ has period N. Then

$$\theta_{u,u}(\tau) = \sum_{t=0}^{N-1} u(t+\tau)u^*(t) = \frac{1}{q-1} \sum_{t=0}^{q^m-2} (-1)^\tau \chi(s(t+\tau)s(t))$$

Since each pair $(s(t + \tau), s(t)) \neq (0, 0)$ occurs q^{m-2} times in an *m* sequence when *t* runs through $t = 0, 1, \ldots, q^m - 2$, it follows that each nonzero element in GF(q) appears $q^{m-1}(q-1)$ times as a product $s(t + \tau)s(t)$. Hence

$$\theta_{u,u}(\tau) = (-1)^{\tau} q^{m-1} \sum_{x \in \operatorname{GF}(q)} \chi(x) = 0,$$

since the numbers of squares equals the number of nonsquares in $\mathrm{GF}(q)$.

Example 7. This example constructs an Ipatov sequence of period 13 from an *m* sequence of period $n = 3^3 - 1 = 26$. Note that $\chi(0) = 0, \chi(1) = 1, \chi(2) = -1$. Let + and - denote +1 and -1 respectively. The *m* sequence

00202122102220010121120111

leads to the following ternary Ipatov sequence of period 13

$$00 - 0 - - - + + 0 - + -$$

which possesses perfect autocorrelation.

Ipatov sequences can be constructed for all length $n = (q^m - 1)/(q - 1)$ whenever p and m are odd. Høholdt and Justesen have constructed ternary sequences with perfect autocorrelation also in the case when $q = 2^r$ is even.

7. THE CROSS-CORRELATION OF *m* SEQUENCES

In this section we study the cross-correlation between two different *m* sequences of period $n = p^m - 1$ with symbols from GF(p), where *p* is a prime. Many families of sequences with good correlation properties can be derived from these cross-correlation properties.

Any *m* sequence of length $n = p^m - 1$ can (after a cyclic shift) be obtained by decimating $\{s(t)\}$ by a *d* relatively prime to *n*. We use $C_d(\tau)$ to denote the cross-correlation function between the *m* sequence $\{s(t)\}$ and its decimation $\{s(dt)\}$. By definition, we have

$$C_d(\tau) = \sum_{t=0}^{p^m - 2} \omega^{s(t+\tau) - s(dt)}$$

We can assume after suitable shifting that $\{s(t)\}$ is in the form $s(t) = \text{Tr}(\alpha^t)$, where Tr(x) denotes the trace function from $GF(p^m)$ to GF(p). Using the properties of the trace function, this can be reformulated as

$$\begin{split} C_d(\tau) &= \sum_{t=0}^{p^m-2} \omega^{\operatorname{Tr}(\alpha^{t+\tau}) - \operatorname{Tr}(\alpha^{dt})} \\ &= \sum_{x \in \operatorname{GF}(p^m) \setminus \{0\}} \omega^{\operatorname{Tr}(cx-x^d)} \end{split}$$

where $c = \alpha^{\tau}$.

Hence, finding the cross-correlation function between *m* sequences is equivalent to evaluating the sum above. Such sums, called *exponential sums*, have been extensively studied in the literature. Several of the best sequence families applied in CDMA systems are based on estimates of such exponential sums.

In the case when $\{s(t)\} = \{s(dt)\}\)$, that is, when $d \equiv p^i \pmod{p^m - 1}$, then we have the two-valued autocorrelation function of the *m* sequence $\{s(t)\}\)$. When the two sequences are cyclically distinct, that is, when $d \neq p^i \pmod{p^m - 1}$, then the cross-correlation function $C_d(\tau)$ is known to take on at least three different values when $\tau = 0, 1, \ldots, p^m - 2$.

It is therefore of special interest to study the cases when exactly three values occur. In addition to being a long studied and challenging mathematical problem, it turns out that several of these cases often lead to a low maximum absolute value of the cross-correlation. In the binary case when p = 2, the following six decimations give three-valued cross-correlation; these decimations cover all the known binary cases, but there are no proofs that these are the only ones:

1.
$$d = 2^k + 1, m/\gcd(m, k)$$
 odd.
2. $d = 2^{2k} - 2^k + 1, m/\gcd(m, k)$ odd.
3. $d = 2^{m/2} + 2^{(m+2)/4} + 1, m \equiv 2 \pmod{4}$.
4. $d = 2^{(m+2)/2} + 3, m \equiv 2 \pmod{4}$.
5. $d = 2^{(m-1)/2} + 3, m$ odd.
6. $d = \begin{cases} 2^{(m-1)/2} + 2^{(m-1)/4} - 1 & \text{if } m \equiv 1 \pmod{2^{(m-1)/2} + 2^{(3m-1)/4} - 1} & \text{if } m \equiv 3 \pmod{4} \end{cases}$

Case 1 is the oldest result due to Gold [4] in 1968, and forms the basis for the binary Gold sequences found in numerous applications. Case 2 produces sequences with properties similar to those of the Gold sequences and was first proved by Kasami and Welch in the late 1960s ago.

4) 4)

In the nonbinary case when p > 2, fewer cases give three-valued cross-correlation. The following two cases have a three-valued cross-correlation; these were proved by Trachtenberg [3] for *m* odd and extended to the case m/gcd(m, k) odd in Helleseth [8]:

1.
$$d = (p^{2k} + 1)/2, m/\gcd(m, k)$$
 odd.
2. $d = p^{2k} - p^k + 1, m/\gcd(m, k)$ odd.

These decimations are analogs to the Gold and Kasami–Welch decimations, respectively. The three values of the cross-correlation that occur in these two cases are $\{-1, -1 \pm p^{(m+e)/2}\}$, where $e = \gcd(k, m)$. The maximum

absolute values of the cross-correlation function are smallest in the case when *m* is odd and gcd(k,m) = 1, when the cross-correlation values are $\{-1, -1 \pm p^{(m+1)/2}\}$. For p > 3, there are no other decimations known that have a three-valued cross-correlation.

Dobbertin et al. [9] found new decimations for ternary m sequences that give three-valued cross-correlation. The family of ternary sequences presented below do not have a known analogue when p > 3.

Theorem 4. Let $d = 2 \cdot 3^{(m-1)/2} + 1$, where *m* is odd, then the cross-correlation function $C_d(\tau)$ takes on the following three values:

$$\begin{array}{ccc} -1+3^{(m+1)/2} & \text{occurs} & \frac{1}{2}(3^{m-1}+3^{(m-1)/2}) & \text{times} \\ -1 & \text{occurs} & 3^m-3^{m-1}-1 & \text{times} \\ -1-3^{(m+1)/2} & \text{occurs} & \frac{1}{2}(3^{m-1}-3^{(m-1)/2}) & \text{times} \end{array}$$

Numerical results have also revealed some other decimations with the same cross-correlation properties. At present this is the only known open case of three-valued cross-correlation of m sequences.

Conjecture 1. Let $d = 2 \cdot 3^r + 1$, where *m* is odd, and

$$r = \begin{cases} \frac{m-1}{4} & \text{if } m \equiv 1 \pmod{4} \\ \frac{3m-1}{4} & \text{if } m \equiv 3 \pmod{4} \end{cases}$$

then the cross-correlation function $C_d(\tau)$ is as in Theorem 4.

New ternary sequences of period $n = 3^m - 1$, with the same autocorrelation values as *m* sequences, have been constructed by adding two *m* sequences of this period. Two simple examples are given below. Let $d = 2 \cdot 3^{(m-1)/2} + 1$, where *m* is odd or $d = 2^{2k} - 2^k + 1$, where m = 3k [10]. Let $s(t) = \text{Tr}(\alpha^t)$; then the ternary sequence $\{u(t)\}$, where

$$u(t) = s(t) + s(dt)$$

has the same two-level autocorrelation function as the m sequence of the same period. More recently, these results have been generalized to nonternary sequences.

8. BOUNDS ON SEQUENCE CORRELATIONS

Practical applications require a family \mathcal{F} of sequences of period n, such that the number of users $M = |\mathcal{F}|$ is large and simultaneously the maximum value of the autocorrelation and crosscorrelation, C_{\max} , is as small as possible. Welch and Sidelnikov provide important lower bounds on the minimum value of C_{\max} for a family of sequences of size M and period n. These bounds are important in comparing different sequence designs with the optimal possible achievable parameters.

The following bound is due to Welch [2].

Theorem 5. Let \mathcal{F} be a family of M complex-valued sequences of period n

$$\mathcal{F} = \{\{b_i(t)\}: i = 1, 2, \dots, M\}$$

where each sequence has norm (or energy) *n*:

$$\sum_{t=0}^{n-1} |b_i(t)|^2 = n, \quad \text{for all } i = 1, 2, \dots, M$$

Let ρ_{\max} be the maximum nontrivial correlation value of the family $\mathcal F$

$$ho_{\max} = \max \left\{ egin{array}{c} \sum_{t=0}^{n-1} b_i(t+ au) b_j^*(t) \colon 1 \leq i, \ j \leq M ext{ either } i
eq j ext{ or } au
eq 0
ight.
ight.$$

where * denotes complex conjugation. Then for all $k \ge 0$, it holds that

$$ho_{ ext{max}}^{2k} \geq rac{1}{nM-1} \left\{ rac{Mn^{2k+1}}{inom{k+n-1}{n-1}} - n^{2k}
ight\}$$

To obtain a lower bound on the maximum nontrivial correlation for a family of sequences, $\{b_i(t)\}$, over an alphabet of size p, one applies the Welch bound to the family

$$\mathcal{F} = \{\{\omega^{b_i(t)}\}: i = 1, 2, \dots, M\}$$

The following bound is due to Sidelnikov [1].

Theorem 6. Let \mathcal{F} be a family of M sequences of period n over an alphabet of size p. In the case p = 2, then

$$egin{aligned} C_{ ext{max}}^2 &> (2k+1)(n-k) + rac{k(k+1)}{2} \ &-rac{2^k n^{2k+1}}{m(2k)! \left(egin{aligned} n \ k \end{array}
ight)}, \quad 0 \leq k < rac{2n}{5} \end{aligned}$$

In the case p > 2, then

$$C_{\max}^2 > \left(rac{k+1}{2}
ight)(2n-k) - rac{2^k n^{2k+1}}{m(k!)^2 \left(rac{2n}{k}
ight)}, \quad k \ge 0$$

The Welch bound applies to complex-valued sequences, while the Sidelnikov bound applies to sequences over any finite alphabet. To find the best result one normally has to examine both bounds. Some improvements of these bounds have been obtained by Levenshtein [11]. Applying the Sidelnikov bound for a family of size $M = n^u$ for some $u \ge 1$, one observes that the best bound is usually obtained by setting $k = \lfloor u \rfloor$. The Sidelnikov bound is well approximated, when $n \gg u$, by

 $C_{\max}^2 > n \left\{ (2u+1) - \frac{1}{(2u-1)!!} \right\}$ for p = 2

and

$$C_{\max}^2 > n \left\{ (u+1) - \frac{1}{(u)!} \right\}$$
 for $p > 2$

where (2u-1)!! denotes $1 \cdot 3 \cdot 5 \cdots (2u-1)$. From these approximations, one sees that the bounds indicate that one may improve the performance in terms of C_{\max} using

a nonbinary alphabet instead of a binary alphabet. The improvement can be obtained for a small alphabet, even for ternary sequences.

Using the tightest lower bound on the maximum correlation parameters obtained from the Welch or Sidelnikov bounds one can derive asymptotic bounds. For example, in the case when the size of the sequence family $M = |\mathcal{F}|$ is approximately equal to the length *n*, the lower bounds are $C_{\max} \approx (2n)^{1/2}$ for p = 2 and $C_{\max} \approx n^{1/2}$ for p > 2.

9. SEQUENCE FAMILIES WITH LOW CORRELATIONS

This section surveys and compares some of the best-known ternary sequence designs for CDMA applications. Several of the sequence designs are asymptotically optimal and give good parameters also when $p \neq 3$. First we describe sequences based on the crosscorrelation of *m* sequences.

Theorem 7. Let $\{s(t)\}$ and $\{s(dt)\}$ be two *m* sequences of period $n = p^m - 1$, *p* a prime. Let *S* be the family of sequences given by

$$\begin{split} \mathcal{S} &= \{\{s(t+\tau) + s(dt)\}: \tau = 0, 1, \dots, p^m - 2\} \\ &\cup \{s(t)\} \cup \{s(dt)\} \end{split}$$

Then $|\mathcal{S}| = n + 2 = p^m + 1$ and the maximum nontrivial auto- or cross-correlation value in the family is $C_{\max} = \max\{|C_d(\tau)|: \tau = 0, 1, \dots, n-1\}.$

Proof The idea behind this construction is that the cross-correlation between any two sequences in S equals the cross-correlation between two *m* sequences that differ by a decimation *d*. Let $\{u_1(t)\}$ and $\{u_2(t)\}$ be two sequences in S. Consider the typical case when $u_1(t) = s(t + \tau_1) + s(dt)$ and $u_2(t) = s(t + \tau_2) + s(dt)$.

Using the shift-add property of *m* sequences, the difference $u_1(t + \tau) - u_2(t)$ can be calculated by

$$u_{1}(t + \tau) - u_{2}(t) = (s(t + \tau_{1} + \tau) + s(d(t + \tau)))$$
$$- (s(t + \tau_{2}) + s(d(t)))$$
$$= (s(t + \tau_{1} + \tau) - s(t + \tau_{2}))$$
$$- (s(dt) - s(dt + d\tau))$$
$$= s(t + \gamma_{1}) - s(dt + \gamma_{2})$$
$$= s(t + \gamma) - s(dt)$$

Therefore, the correlation between two sequences in S equals the cross-correlation between $\{s(t)\}$ and $\{s(dt)\}$, which has a maximal absolute value $C_{\max} = \max\{|C_d(\tau)|: \tau = 0, 1, \dots, n-1\}.$

It follows from this result and the known results on the cross-correlation function of m sequences that the best sequence designs of this form are obtained for the values of d for which the maximum absolute value of the crosscorrelation is as small as possible. This happens in the following cases:

1. $d = (p^{2k} + 1)/2$, m odd, gcd(k, m) = 1. 2. $d = p^{2k} - p^k + 1$, m odd, gcd(k, m) = 1. 3. $d = 2 \cdot 3^{(m-1)/2} + 1$, m odd and p = 3. These cases give $C_{\max} = p^{(m+1)/2} + 1 = (p(n+1))^{1/2} + 1$, and thus the parameters of the sequence families above are

$$M = |S| = n + 1, n = p^m - 1, \text{ and } C_{\max} = (p(n + 1))^{1/2} + 1$$

For m even, there are better constructions using the cross-correlation of other pairs of m sequences. The best sequence family of this type has parameters

$$M = |S| = n + 1, n = p^m - 1, \text{ and } C_{\max} = 2(n + 1)^{1/2} - 1$$

This family is based on *m* sequences with a fourvalued cross-correlation function with values $\{-1 - p^{m/2}, -1, -1 + p^{m/2}, -1 + 2p^{m/2}\}$; see Helleseth [8], where

$$d = 2p^{m/2} - 1 \text{ and } p^{m/2} \neq 2 \pmod{3}$$

It is natural to compare the parameters of these designs with the best possible that may be achieved according to the Welch or Sidelnikov bound. For a sequence family of size approximately equal to the length (i.e., $M \approx n$), both bounds give an asymptotic lower bound on $C_{\max} \approx n^{1/2}$ (for p > 2). The best sequence families obtained from Theorem 7 are a factor $p^{1/2}$ off the optimal value for m odd and a factor of 2 off the optimal value for m even.

It is interesting to observe that the popular binary Gold sequence family has also size $M \approx n$, and $C_{\max} \approx (2n)^{1/2}$. According to the Sidelnikov bound for p = 2, this is optimal for binary sequences. The Welch and Sidelnikov bounds indicate that this can be improved with a factor of $2^{1/2}$ for nonbinary sequences since asymptotically $C_{\max} \approx n^{1/2}$ may be achievable.

We next construct sequence families better than the ones obtained from the cross-correlation of m sequences. These sequence families are asymptotically optimal with respect to the Welch and Sidelnikov bounds. These sequences can be described using *perfect nonlinear* (*PN*) mappings.

Let f(x) be a function $f: \operatorname{GF}(p^m) \to \operatorname{GF}(p^m)$. Then f is said to be a perfect nonlinear mapping, if f(x + a) - f(x) is a permutation of $\operatorname{GF}(p^m)$ for any nonzero $a \in \operatorname{GF}(p^m)$. From PN mappings with $f(x) = x^d$, one can construct families of sequences with good correlation properties when p is an odd prime.

Example 8. There are three known families of PN mapping of the form $f(x) = x^d$; the first two work for any odd prime *p*, while the third works only for p = 3:

(a) d = 2.
(b) d = p^k + 1 where m/gcd (k, m) is odd.
(c) d = (3^k + 1)/2 where k is odd and gcd (k, m) = 1.

Let ω be a complex *p*th root of unity, and let Tr (*x*) denote the trace mapping from $GF(p^m)$ to GF(p). For any PN mapping, it holds for any nonzero $a \in GF(p^m)$, that

$$\sum_{\operatorname{GF}(p^m)} \omega^{\operatorname{Tr}(f(x+a)-f(x))} = \sum_{b \in \operatorname{GF}(p^m)} \omega^{\operatorname{Tr}(b)} = 0$$

since the trace function takes on all values in $\mathrm{GF}(p)$ equally often.

x∈
Let $c, \lambda \in GF(p^m)$ and $c \neq 0$; then the following exponential sum is the key to the asymptotically optimal sequence families constructed from PN mappings:

$$S(c, \lambda) = \sum_{x \in \operatorname{GF}(p^m)} \omega^{\operatorname{Tr}(cf(x) + \lambda x)}$$

It is important to determine $|S(c, \lambda)|$, which can be done as follows

$$\begin{split} |S(c,\lambda)|^2 &= \sum_{x,y \in \mathbf{GF}(p^m)} \omega^{\operatorname{Tr}(c(f(x)-f(y))+\lambda(x-y))} \\ &= \sum_{y,z \in \mathbf{GF}(p^m)} \omega^{\operatorname{Tr}(c(f(y+z)-f(y))+\lambda z)} \\ &= p^m \end{split}$$

since the PN property only gives a contribution when z = 0. Hence

$$|S(c,\lambda)| = |\sum_{x \in \operatorname{GF}(p^m)} \omega^{\operatorname{Tr}(cf(x) + \lambda x)}| = p^{m/2}$$

This leads to the following asymptotically optimal sequence family.

Theorem 8. Let $f(x) = x^d$ be a PN mapping of GF (p^m) . Let α be a primitive element in GF (p^m) . Let $\{s_c(t)\}$ be the sequence of period $n = p^m - 1$ defined by

$$s_c(t) = \operatorname{Tr}(\alpha^t + cf(\alpha^t))$$

Let \mathcal{F} be the family of sequences

$$\mathcal{F} = \{\{s_c(t)\}: c \in \operatorname{GF}(p^m)\}$$

Then \mathcal{F} is a family of $M = p^m$ sequences of period $n = p^m - 1$ and $C_{\max} \leq 1 + p^{m/2}$.

Proof The cross-correlation between two sequences in the family is

$$\begin{aligned} \theta s_{c_1}, s_{c_2}(\tau) &= \sum_{t=0}^{p^{-2}} \omega^{s_{c_1}(t+\tau) - s_{c_2}(t)} \\ &= \sum_{t=0}^{p^m - 2} \omega^{\operatorname{Tr}((c_1 \alpha^{d\tau} - c_2) \alpha^{dt} + (\alpha^{\tau} - 1) \alpha^t)} \\ &= -1 + \sum_{x \in GF(p^m)} \omega^{\operatorname{Tr}(cf(x) + \lambda x)} \end{aligned}$$

where $c = c_1 \alpha^{d\tau} - c_2$ and $\lambda = \alpha^{\tau} - 1$. Hence, the sum above gives $|\theta s_{c_1}, s_{c_2}(\tau)| \le 1 + p^{m/2}$, except when $c_1 = c_2$ and $\tau = 0$.

Thus, the family \mathcal{F} is asymptotically optimal with respect to the Welch–Sidelnikov bound. The family corresponding to the PN mapping with d = 2 is due to Sidelnikov, while $d = p^k + 1, m/\gcd(k, m)$ odd is due to Kumar and Moreno [6]. The connection between PN functions and these optimal families were pointed out by Helleseth and Sandberg [7], who also described the ternary sequence family for $d = (3^k + 1)/2$, where k is odd and $\gcd(k.m) = 1$. **Example 9.** In the following example we describe a family of ternary Kumar-Moreno sequences of period $n = 3^3 - 1 = 26$ for $d = 3^k + 1$, where k = 1, namely, $d = 3^1 + 1 = 4$. The *i*th sequence in the family is defined to be

$$s_i(t) = \operatorname{Tr}(\alpha^t + \beta_i \alpha^{dt}), \text{ where } \beta_i \in \operatorname{GF}(3^3)$$

Let $u(t) = \operatorname{Tr}(\alpha^t), v_0(t) = \operatorname{Tr}(\alpha^{4t}), v_1(t) = \operatorname{Tr}(\alpha^{4t+1})$, then

u(t) = 00202122102220010121120111 $v_0(t) = 02120112220200212011222020$ $v_1(t) = 01001210221110100121022111$

All sequences in the family \mathcal{F} can be obtained by adding different cyclic shifts of the sequences $\{v_0(t)\}$ or $\{v_1(t)\}$ (both of period 13) to the sequence $\{u(t)\}$. Figure 1 shows the shiftregister that generates all the sequences in \mathcal{F} . The maximal absolute value of the correlation for this family is $C_{\max} \leq 1 + (27)^{1/2}$.

Most of the families discussed above have family size M approximately equal to the length n. Because of the large increase in the number of users in modernday communication systems, there is a demand for larger sequence families. A general construction due to Sidelnikov [1] based on general bounds on exponential sums is presented below. These families have a flexible and larger family size.

The main idea is that if f(x) is a polynomial of degree $d \ge 1$ with coefficients from GF (p^m) that cannot be expressed in the form $f(x) = g(x)^p - g(x) + c$ for any g(x) with coefficients from GF (p^m) and any $c \in GF(p^m)$, then

$$\sum_{x \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(f(x))} \bigg| \le (d-1)\sqrt{p^m}$$

Theorem 9. Let p be a prime, $q = p^m$ and α an element of order $n \mid q - 1$ in GF (q). Consider the family of all sequences $\{s(t)\}$ over GF (p) of the form

$$s(t) = \operatorname{Tr}\left\{\sum_{k=1}^{d} a_k \alpha^{kt}\right\}$$

where $(a_1, a_2, \ldots, a_d) \in GF(q)^d$ and d is an integer satisfying $1 \le d \le p^{\lfloor (m+1)/2 \rfloor}$. Let \mathcal{F} be the subset of this family consisting of those sequences having period equal to n. If M denotes the number of cyclically distinct sequences in this family and C_{\max} the maximum correlation value, then

$$M \geq \frac{q-1}{n} q^{d - \lfloor d/p \rfloor - 1}$$

and

$$C_{\max} \leq \left(rac{d-n}{(q-1)}
ight) q^{1/2} + rac{n}{q-1}$$

The main idea behind this construction is that if $\{s_1(t)\}$ and $\{s_2(t)\}$ correspond to polynomials $f_1(x)$ and $f_2(x)$ of degree $\leq d$, respectively, then the difference



Figure 1. A ternary Kumar–Moreno sequence family.

 $s_1(t + \tau) - s_2(t)$ involved in the correlation calculations corresponds to the polynomial $f_1(cx) - f_2(x)$, where $c = \alpha^{\tau}$. Since $f_1(cx) - f_2(x)$ has degree $\leq d$, the exponential sum bound above gives the result.

The entry attributed to Sidelnikov in Table 3 corresponds to the sequences above with n = q - 1 and d = 2. These sequences are the same as the ones obtained from the PN mapping with d = 2 above.

Bent sequences is another family of sequences that in addition to low cross-correlation have large linear span. This is a family of sequences that exists for all primes p. The sequences in the family have period $n = p^m - 1$ for even values of m. The sequences in the family are balanced and have maximum correlation $C_{\max} = p^{m/2} + 1$. For further information on nonbinary bent sequences, see Kumar et al. [12].

Table 3 compares some of the families of sequences constructed in this section. The families have size M approximately equal to the period n.

For further information on sequences and their correlation properties, the reader is referred to surveys on sequences that can be found in Refs. 13-15.

BIOGRAPHY

Tor Helleseth received the Cand.Real. and Dr.Philos. degrees in mathematics from the University of Bergen, Bergen, Norway, in 1971 and 1979, respectively.

From 1973 to 1980 he was a Research Assistant at the Department of Mathematics, University of Bergen. From 1981 to 1984 he was a Researcher at the Chief Headquarters of Defense in Norway. Since 1984 he has been a Professor at the Department of Informatics at the University of Bergen.

From 1991 to 1993 he served as an Associate Editor for Coding Theory for *IEEE Transactions on Information Theory*. Since 1996 he has been on the editorial board of *Designs, Codes and Cryptography*. He was Program Chairman for Eurocrypt'93 and for the 1997 Information Theory Workshop. He was one of the organizers of the conferences on Sequences and Their Applications SETA'98 and SETA'01. He has published more than 100 scientific papers in international journals.

In 1997 he was elected an IEEE Fellow for his contributions to coding theory and cryptography. His

Family	Length	Alphabet	Family Size	C_{\max}
Trachtenberg [3]	$n = p^m - 1$			
	p prime, m odd	p	n+2	$1 + ((n+1)p)^{1/2}$
Dobbertin et al. [9]	$n = 3^m - 1$			
	$m \operatorname{odd}$	3	n+2	$1 + ((n+1)3)^{1/2}$
Helleseth [8]	$n = p^m - 1$			
	p prime, m even	p	n+2	$-1+2(n+1)^{1/2}$
	$p^{m/2} eq 2 \pmod{3}$			
Sidelnikov [1,5]	p^m-1			
	p prime	p	n+1	$1 + (n+1)^{1/2}$
Kumar–Moreno [6]	$n = p^m - 1$			
	p prime	p	n+1	$1 + (n+1)^{1/2}$
Helleseth–Sandberg [7]	$n = 3^m - 1$	3	n+1	$1 + (n+1)^{1/2}$
Bent sequences	p^m-1			
	p prime, m even	p	$(n+1)^{1/2}$	$1 + (n+1)^{1/2}$

research interests include coding theory, sequence designs, and cryptography.

BIBLIOGRAPHY

- V. M. Sidelnikov, On mutual correlation of sequences, Soviet Math. Doklad. 12: 197–201 (1971).
- L. R. Welch, Lower bounds on the maximum cross correlation of signals, *IEEE Trans. Inform. Theory* **IT-20**: 397–399 (1974).
- H. M. Trachtenberg, On the Cross-Correlation Functions of Maximal Linear Recurring Sequences, Ph.D. thesis, Univ. Southern California, 1970.
- R. Gold, Maximal recursive sequences with 3-valued recursive cross-correlation functions, *IEEE Trans. Inform. Theory* IT-14: 154–156 (1968).
- V. M. Sidelnikov, Cross correlation of sequences, *Probl. Kybern.* 24: 15-42 (1971) (in Russian).
- 6. P. Kumar and O. Moreno, Prime-phase sequences with periodic correlation properties better than binary sequences, *IEEE Trans. Inform. Theory* **IT-37**: 603-616 (1991).
- T. Helleseth and D. Sandberg, Some power mappings with low differential uniformity, Appl. Algebra Eng. Commun. Comput. 8: 363–370 (1997).
- T. Helleseth, Some results about the cross-correlation function between two maximal linear sequences, *Discrete Math.* 16: 209–232 (1976).
- H. Dobbertin, T. Helleseth, V. Kumar, and H. Martinsen, Ternary m-sequences with three-valued crosscorrelation function: Two new decimations, *IEEE Trans. Inform. Theory* IT-47: 1473-1481 (2001).
- T. Helleseth, P. V. Kumar, and H. Martinsen, A new family of sequences with ideal two-level autocorrelation function, *Designs, Codes Cryptogr.* 23: 157-166 (2001).
- V. I. Levenshtein, Bounds on the maximal cardinality of a code with bounded modules of the inner product, *Soviet Math. Doklad.* 25: 526–531 (1982).
- P. V. Kumar, R. A. Scholtz, and L. R. Welch, Bent function sequences, *IEEE Trans. Inform. Theory* **IT-28**: 858–864 (1982).
- J. Gibson, ed., The Mobile Communications Handbook, CRC Press and IEEE Press, New York, 1996.
- 14. V. S. Pless and W. C. Huffman, eds., *The Handbook of Coding Theory*, North-Holland, Amsterdam, 1998.
- M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, Spread-Spectrum Communications Handbook, McGraw-Hill, New York, 1994.

TERRESTRIAL DIGITAL TELEVISION

YIYAN WU Communications Research Centre Canada Ottawa, Ontario, Canada

1. INTRODUCTION

A *terrestrial digital television* (TDT) system broadcasts video, audio, and ancillary data services over the VHF and

UHF bands (50–800 MHz). It is a wideband digital pointto-multipoint transmission system — a high-speed digital pipe for multimedia services to the general public. The payload data rate for a TDT system is between 19 and 25 Mbps (megabits per second). The TDT will eventually replace existing analog television services, such as NTSC, PAL, and SECAM.

The first commercial TDT services were introduced in the United States and in the United Kingdom in November 1998. Since then, Australia, Sweden, Spain, Singapore, and Korea have also started TDT broadcast services. Many countries are having field trials and pilot projects of TDT systems and are planning to launch commercial service in the near future. Existing analog TV services are expected to be phased out by 2010–2020.

The bandwidth of a TDT system can either be 6, 7, or 8 MHz, depending on the countries and regions. Generally, the TDT system bandwidth is identical to the bandwidth of the analog television system that it will replace in any particular country.

The advantages of a TDT system in comparison to an analog television system are typically

- 1. Better Picture Quality. A TDT system can deliver high-definition television (HDTV) pictures at about twice the vertical resolution of an analog television system with the same bandwidth. It can also provide wide-screen formatted pictures with an aspect ratio of 16:9 rather than the traditional near-square, 4:3 television format.
- 2. Better Audio Quality. A TDT system can deliver CD-quality stereo or multichannel audio (5.1 channels). It can also carry multiple audio channels for multilanguage implementation.
- 3. *Ghost-Free and Distortion-Free Pictures*. There are no ghosts or any other forms of transmission distortions (e.g., co- and adjacent-channel interference, tone interference, impulse noise) visible on the screen.
- 4. *High Spectrum and Power Efficiency*. Each TDT channel can carry up to eight analog TV quality programs. In addition, the TDT system requires much less transmission power for the same coverage.
- 5. User-Friendly Interface. The TDT system provides an electronics program guide (EPG) for easy program and channel selection. Time-shifted viewing can easily be implemented by preselecting programs for recording.
- 6. *Easy Interface with Other Media*. Since a TDT system is a high-speed "digital pipe," it can have seamless interfaces with other communication systems and computer networks. Nonlinear editing, lossless storage, and access to video databases can easily be implemented.
- 7. *Multimedia Services and Data Broadcasting*. As a fully digital system, the TDT system can be used to deliver multimedia services and to provide data broadcasting service. Interactive services will also be available.

2548 TERRESTRIAL DIGITAL TELEVISION

- 8. *Conditional Access*. In a TDT system, it is possible to address each receiver for content control or for value-added services.
- 9. Providing Fixed, Mobile, or Portable Services. Based on reception conditions, the TDT system can provide high-data-rate service for fixed reception, or intermediate-data-rate service for portable reception, or lower-data-rate service to automobilemounted high-speed mobile reception.
- 10. Single-Frequency Network (SFN) or Diversified Transmission. Because of the strong multipath immunity of the TDT system, it is possible to operate a series of TDT transmitters on the same RF frequency to provide better coverage and service quality, and to achieve better spectrum efficiency.

The disadvantages of a TDT system in comparison to an analog television system are

- 1. Sudden Service Dropout. This is the "cliff effect." As a digital transmission system, the TDT system video and audio signals fail abruptly when the signal-to-noise ratio (SNR) drops below a critical operating threshold. Analog TV systems, on the other hand, have graceful degradation of signal power versus picture quality and the audio is extremely robust.
- 2. Longer Acquisition Time. For an analog TV system, signal acquisition is almost instantaneous. A TDT system typically takes about half a second (0.5 s) to acquire a signal. This can be bothersome when changing channels or channel surfing. The delay is mostly caused by the video decoding system. Synchronization, channel estimation, equalization, channel decoding, and conditional access also contribute to the delay.
- 3. *Frame Rate.* The TDT system uses the same frame rate as the analog TV system: 25 or 30 frames per second. For high-intensity pictures, the screen can show an annoying flicking, especially for large-screen-display devices.

One of the most challenging issues of introducing the TDT is that the TDT service has to coexist for some time in the same spectrum band with the existing analog TV service in a frequency-interlaced approach. Spectrum is a valuable and limited resource. In most countries and regions, there is no additional spectrum in the VHF/UHF bands available for TDT implementation. The TDT system has to use "taboo channels" [1], those channels that are unusable for broadcasting analog TV services, because of the existence of intermodulation products and other forms of interference, such as co-channel and adjacent channel interference. In other words, a TDT system has to be robust to various types of interference. The analog TV system is quite sensitive to interference (even if the interference level is 50 dB below the analog TV signal, viewers can still notice it). The service quality of the analog TV should not suffer from much degradation due to the introduction of the TDT service. Therefore, the TDT system has to transmit at low power, or, in other words, to operate at low SNR, which means that strong channel coding and forward error correction (FEC) techniques have to be implemented.

Video compression, or video source coding, is another challenge. The raw data rate of a HDTV program (without data compression) is about 1 Gbps. The TDT system data throughput is only 18-22 Mbps. Therefore, the video source coding system has to achieve at least a 50:1 data compression ratio, in real time, and still provide a high video quality.

2. SYSTEM DESCRIPTION

As shown in Fig. 1, a TDT system typically consists of source coding, multiplexing, and transmission systems. The video and audio source coding are implemented to compress the video and audio source data to reduce the data rate. The compressed data are multiplexed with the program-related data, such as the electronics program guide (EPG), the Program and System Information Protocol (PSIP) or services information (SI) data, program rating data, closed captioning data, and conditional access data, as well as other data that are not related to the television programs, such as opportunistic data and bandwidth-guaranteed data. The multiplexed output, or transport stream (TS), is channel coded and modulated for transmission over a terrestrial RF channel.

At the receiving end, the signal undergoes demodulation and FEC to correct the transmission errors. The resulting data are demultiplexed to sort out video output data, audio output data, and other program-related and non-program-related data.

2.1. Transmission System

2.1.1. Transmission System Description

2.1.1.1. Channel Coding. Figure 2 is a diagram for a TDT transmission system. As mentioned earlier, the combination of low power emission and robustness against interference requires that powerful channel coding must be implemented to reduce the TDT system SNR threshold. This is achieved in practice by the use of concatenated channel coding. In such a coding system, two levels of FEC codes are employed: an "inner" modulation or convolution code and an "outer" symbol error correction code. Bandwidth-efficient trellis-coded modulation (TCM) or convolutional coding [2] is implemented as the inner code to achieve high coding gain over the additive white Gaussian noise (AWGN) channel and to correct short bursts of interference, such as those created by analog TV synchronization pulses. This code has good noise-error performance at low SNR. The required output BER for the inner coder is on the order of 10^{-3} to 10^{-4} . The Reed-Solomon (RS) code [3] is used as the outer code to handle the burst of errors generated by the inner code and to provide a system BER of 10^{-11} or lower. The merit of a concatenated coding system is that it can achieve a very low SNR threshold. The drawback is that it has a "cliff" threshold, meaning that the signal dropout is within 1 dB decrease of SNR near the threshold.



Figure 2. Terrestrial digital television transmission system diagram.

In Fig. 2, an interleaver and deinterleaver are used to fully exploit the error correction ability of the FEC code [3]. Since most errors occur in bursts (this is especially true for the outer code), they often exceed the error correction capability of the FEC code. Interleaving is used to spread out, or decorrelate, the burst of errors into shorter error sequences or isolated errors that are within the capability of the FEC code.

2.1.1.2. Modulation. From Fig. 2, the input data, or transport stream data, are coded and interleaved by the outer and inner error correction codes and, then, modulated and frequency shifted to a RF channel for transmission.

There are two types of modulation techniques used in the TDT systems; single-carrier modulation (SCM), such as vestigial sideband (VSB) [4] modulation, and multicarrier modulation (MCM), such as orthogonal frequencydivision multiplexing (OFDM) [5,6] modulation.

In a SCM system, the information-bearing data are used to modulate one carrier, which occupies the entire RF channel. VSB is a one-dimensional modulation scheme. Its symbol rate is about twice the usable bandwidth. The information bits are transmitted in a vestigial sideband, occupying a bandwidth that is slightly larger than half the symbol rate. VSB modulation is mathematically equivalent to the offset quadrature amplitude modulation (OQAM).

Adaptive equalization is implemented in SCM systems to combat the intersymbol interference (ISI) that is usually caused by multipath distortion [4]. A raised-cosine filter is generally used for spectrum shaping. In a MCM system, quadrature amplitude modulation (QAM) [4] is used to modulate multiple low-datarate carriers, which are transmitted concurrently using frequency-division multiplexing (FDM) [4]. Since each QAM carrier spectrum envelope is of the form $\sin(x)/x$, that is, with periodic spectrum zero-crossing points, the QAM carrier spacing can be carefully selected so that each carrier spectrum peak is located on all the other carriers spectrum zero-crossing points. Although the carrier spectra overlap, they do not interfere with each other and the information bits they carry can be demodulated independently without intercarrier interference. In other words, all carriers maintain orthogonality in a FDM fashion. Hence the name OFDM. Spectrum shaping is not needed for OFDM modulation.

The OFDM can be efficiently implemented via the digital fast Fourier transform (FFT), where an inverse FFT is used as the OFDM modulator and a FFT as the demodulator [5,6]. Each inverse FFT output block is called an OFDM symbol, which contains multiple data samples. The FFT sizes used in the TDT systems are 2048, 4096, and 8192. One important feature of the OFDM is the use of a guard interval (GI). By inserting a GI between the OFDM symbols, or FFT blocks, using "cyclic extension" [5,6], the intersymbol interference can be eliminated. However, the amplitude and phase

distortion, or fading, within an OFDM symbol still exists. Channel coding and equalization are used to mitigate the fading channel. Channel coding combined with the OFDM is called coded OFDM or COFDM.

2.1.2. Terrestrial Digital Television Transmission Standard. Currently, there are three TDT transmission standards [7]:

- 1. The Advanced Television Systems Committee (ATSC) system [8] standardized by the ATSC, a United States-based digital television standards body
- 2. The Digital Video Broadcasting Terrestrial (DVB-T) standard [9] developed by the DVB project, a European-based digital broadcasting standard body, and standardized by the European Telecommunication Standard Institution (ETSI)
- 3. The Integrated Service Digital Broadcasting—Terrestrial (ISDB-T) standard [10] developed and standardized by the Association of Radio Industries and Businesses (ARIB) in Japan.

The main characteristics of the three TDT systems are summarized in Table 1. All three systems employ concatenated channel coding. All three standards can be

Systems	ATSC 8-VSB	DVB-T COFDM	ISDB-T BST-OFDM	
Source coding				
Video	Main profile syntax of ISO/IEC 13818-2 (MPEG-2-video)			
Audio	ATSC Standard A/52 (Dolby AC-3) ISO/IEC 13818-3 (MPEG-2—layer II audio) and Dolby AC-3		ISO/IEC 13818-7 (MPEG-2—AAC audio)	
Transport stream	ISO/I	EC 13818-1 (MPEG-2 TS) transport	t stream	
Transmission system Channel coding				
Outer coding	RS (207, 187, $t = 10$) RS (204, 188, $t = 8$)		188, t = 8)	
Outer interleaver	52 RS block interleaver 12 RS block interleaver			
Inner coding	Rate $\frac{2}{3}$ trellis code	Punctured convolutional code — rate: $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, $\frac{5}{6}$, $\frac{7}{8}$ Constraint length = 7, Polynomials (octal) = 171, 133		
Inner interleaver	12:1 trellis-code interleaving	Bitwise interleaving and frequency interleaving	Bitwise interleaving, frequency interleaving, and selectable time interleaving	
Data randomization	16-bit PRBS	16-bit PRBS	16-bit PRBS	
Modulation	8-VSB	COFDM Subcarrier modulation: QPSK, 16 QAM and 64 QAM hierarchical modulation: multiresolution constellation (16 QAM and 64 QAM) Guard intervals: $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$ of OFDM symbol 2 modes: 2K, 8K FFT	BST-OFDM with 13 frequency segments Subcarrier modulation: DQPSK, QPSK, 16 QAM, 64 QAM Hierarchical transmission: choice of three different subcarrier modulations on each segment Guard intervals: $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$ of OFDM symbol 3 modes: 2K, 4K, 8K FFT	

Table 1. Main Characteristics of Three Terrestrial Digital Television Systems

scaled to any RF channel bandwidth (6, 7, or 8 MHz) with corresponding scaling in the data capacity.

It is believed that China is developing another TDT transmission system, which will be finalized in the 2002/2003 timeframe.

2.1.2.1. The ATSC 8-VSB System. The ATSC system [8] uses trellis-coded 8-level vestigial sideband (8-VSB) modulation with the RS (Reed–Solomon) (207, 187) [3] as the outer code and Ungerboeck rate- $\frac{2}{3}$ TCM [2] as the inner code. A raised-cosine filter with a roll-off factor of 11.5% is used for spectrum shaping.

The ATSC system was designed to transmit highquality video and audio (HDTV) and ancillary data over a 6-MHz channel in the VHF/UHF band. The data throughput is 19.4 Mbps. The system was designed to eventually replace the existing analog TV service in the same frequency band. Therefore, one of the system requirements was to allow the allocation of an additional digital transmitter with equivalent coverage for each existing analog TV transmitter. Another requirement was to cause minimum disturbance to the existing analog TV service in terms of service area and population.

Various picture qualities can be achieved using one of 18 possible video formats (standard definition or highdefinition pictures, progressive or interlaced scan format, as well as different frame rates and aspect ratios). The system can accommodate fixed and possibly portable reception.

The system is designed to withstand different types of interference: existing analog TV services, white noise, impulse noise, phase noise, continuous-wave, and multipath distortions. The system is also designed to offer spectrum efficiency and ease of frequency planning.

The main characteristics of the ATSC 8-VSB system are listed in Table 1. It should be mentioned that currently there is an ongoing 8-VSB enhancement project that will enable the ATSC system to accommodate different transmission modes (2-VSB, 4-VSB, and 8-VSB modulation) in order to offer a tradeoff between data rates and system robustness. It could also transmit dual datastreams (a robust datastream and a high-speed datastream time-division multiplexed, i.e., mixed-mode operation with different mixed ratios) within one RF channel. The project is to be completed by mid 2002.

2.1.2.2. DVB-T COFDM System. The DVB-T system [9] uses the coded orthogonal frequency-division multiplexing (COFDM) modulation system with the RS (204, 188) [3] as the outer code and a punctured convolutional code (rates: $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, $\frac{5}{6}$, $\frac{7}{8}$; constraint length = 7; polynomials (octal) = 171, 133) as the inner code. Different QAM modulations (QPSK, 16 QAM, and 64 QAM) can be implemented on the OFDM carriers. The system was designed to operate within the existing UHF spectrum allocated to analog television transmission. The payload data rates range between 4 and 32 Mbps, depending on the choice of channel coding parameters, modulation type, guard interval duration, and channel bandwidth. DVB-T can also accommodate a large range of SNR and different types of channels. It allows fixed, portable, or mobile

reception, with a consequential trade-off in the usable bit rate.

The OFDM system has two operational modes: a "2K mode," which uses a 2048-point FFT, and an "8K mode," which requires an 8192-point FFT. The system makes provisions for selection between different levels of QAM modulation and different inner code rates and also allows two-level hierarchical channel coding and modulation. Moreover, a guard interval with selectable length separates the transmitted symbols, which makes the system robust to multipath distortion and allows the system to support different network configurations, such as large area SFNs and single transmitter operation. The "2K mode" is suitable for single transmitter operation and for small-scale SFN networks. The "8K mode" can be used both for single transmitter operation and large SFN networks.

The main characteristics of the DVB-T COFDM system are listed in Table 1.

2.1.2.3. ISDB-T BST-OFDM. The ISDB-T system [10] uses the band-segmented transmission (BST)-OFDM modulation system. It uses the same channel coding as the DVB-T system, namely, the RS (204, 188) [3] as the outer code and a punctured convolutional code [rate $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, $\frac{5}{6}$, $\frac{7}{8}$; constraint length 7; polynomials (octal) = 171, 133] as the inner code. It also uses a large time-interleaver (≤ 0.5 s) to deal with signal fading in mobile reception and to mitigate impulse noise interference.

The ISDB-T system is intended to deliver digital television, sound programs, and offer multimedia services to fixed, portable, and mobile terminals in the VHF and UHF bands. To meet different service requirements, the ISDB-T system provides a range of modulation and error protection schemes. The payload data rate ranges between 3.7 and 31 Mbps; again, with a consequential tradeoff in the robustness of the transmission.

The system uses a modulation method referred to as *band-segmented transmission* (BST) OFDM, which consists of a set of common basic frequency blocks called BST segments. Each segment has a bandwidth corresponding to $\frac{1}{14}$ th of the terrestrial television channel bandwidth. Thirteen segments are used for data transmission within one terrestrial television channel and one segment is utilised as guard band.

The BST-OFDM modulation provides hierarchical transmission capabilities by using different carrier modulation schemes and coding rates of the inner code on different BST segments. Each data segment can have its own error protection scheme (coding rates of inner code, depth of the time interleaving) and type of modulation (QPSK, DQPSK, 16 QAM, or 64 QAM). Each segment can then meet different service requirements. A number of segments may be combined flexibly to provide a wideband service (e.g., HDTV). By transmitting OFDM segment groups with different transmission parameters, hierarchical transmission is achieved. Up to three service layers (three different segment groups) can be provided in one terrestrial channel. Partial reception of services contained in the transmission channel can be obtained using a narrowband receiver that has a bandwidth as low as one OFDM segment.

Vertical Scanlines	Horizontal Samples per Line	Picture Aspect Ratio	Frame Rate	Scan Format
	C	a. ATSC System		
1080p	1920	16:9	24, 29.97	Progressive
1080i	1920	16:9	29.97	Interlaced
720p	1280	16:9	24, 29.97, 59.94	Progressive
480p	704	4:3, 16:9	24, 29.97, 59.94	Progressive
480i	704	4:3, 16:9	29.97	Interlaced
480p	640	4:3	24, 29.97, 59.94	Progressive
480i	640	4:3	29.97	Interlaced
	b	. ISDB-T System	n	
1080i	1920	16:9	29.97	Interlaced
1080i	1440	16:9	29.97	Interlaced
720p	1280	16:9	59.94	Progressive
480p	720	16:9	59.94	Progressive
480i	720	16:9, 4:3	29.97	Interlaced
480i	544	16:9, 4:3	29.97	Interlaced
480i	480	16:9, 4:3	29.97	Interlaced

Table 2. Video Formats

The main characteristics of the ISDB-T BST-OFDM system are listed in Table 1.

2.2. Source Coding

From Fig. 1, two types of source coding are implemented in a TDT system: video coding and audio coding.

2.2.1. Video Coding and Video Formats. All TDT systems adopted the MPEG-2, or ISO/IEC 13818-2, video compression standard [11]. ISO/IEC 13818-2 is a discrete cosine transform (DCT) based, motion-compensated interframe coding. It was developed by the Motion Picture Experts Group (MPEG). The standard supports a wide range of picture qualities, data rates, and video formats for broadcast and multimedia applications (see Table 2).

Different video formats are used by the TDT systems from different parts of the world. In Europe, TDT is implemented to broadcast multiple Standard Definition Television (SDTV) signals over a TDT RF channel, which is called a "multiplexer." The video format is the same as the analog TV system in Europe, specifically, 720 pixels per scanline, 576 interlaced active scanlines per frame, and 50 frames per second, or a 50-Hz 720 \times 576*i* format. The aspect ratios are 4:3 and 16:9. The letter "i" indicates interlaced scanning, a scheme which displays every second scanline and then fills in the gaps in the next pass. "Interlacing" is a scanning format used in the analog TV system to reduce the TV signal bandwidth by one-half [1]. The impact of using interlaced scanning is a slight reduction of vertical resolution.

In North America, the TDT system was designed to accommodate both HDTV and SDTV formats. Table 2, part (a) lists the 18 video formats supported by the ATSC standard [8]. Although this table is not mandated by the U.S. Federal Communications Commission (FCC), it is supported by all consumer receiver manufacturers as a voluntary industry standard. In Japan, the TDT system was also planned to provide the HDTV and SDTV services. Table 2, part (b) shows the available video formats.

Currently, two HDTV formats are used in TDT broadcasting: $30 \text{ Hz} - 1920 \times 1080i$ and $60 \text{ Hz} - 1280 \times 720p$, where "p" stands for the progressive (noninterlaced) scan format, which is widely used on computer terminals. For comparable picture quality, the progressive scanning format has twice the frame rate but requires fewer scanlines. Both HDTV formats use a 16:9 aspect ratio.

2.2.2. Audio Coding. Three audio coding standards are implemented in the TDT systems:

- 1. MPEG layer II audio coding
- 2. Dolby AC-3 multichannel audio system
- 3. MPEG-2 advanced audio coding (AAC)

The MPEG audio coding was developed by the Motion Picture Experts Group (MPEG). It can further be classified into MPEG-1 audio coding, or ISO/IEC 11172-3 [12]; and MPEG-2 audio coding, or ISO/IEC 13818-3 and ISO/IEC 13818-7 [11].

The MPEG-1 audio coding has three layers: I, II, and III. Layers I and II are based on subband coding using a 32-subband filterbank. The coded information includes scale factors per band, the bit allocation for each band, and the quantized values for each band. The MPEG-1 layer III is based on the same subband filterbank followed by a modified discrete cosine transform (MDCT) stage that yields a filterbank with 576 bands. The quantized values are Huffman-coded to obtain greater efficiency. The MPEG-1 layer III is often called MP3 for downloading compressed music files via the Internet.

The MPEG-2 audio (ISO/IEC 13818-3 [11]) specifications extend the MPEG-1 layer II coders to lower sample rates and to multichannel audio. Dolby AC-3 [13,14] is based on an MDCT filterbank with 256 bands. A differentially coded representation of the spectral envelope is transmitted along with the quantized values. The bit allocation is derived from the spectral envelope, identically by the encoder and decoder, with the encoder having the ability to adjust to the psychoacoustic model. The Dolby AC-3 multichannel audio system can provide the 5.1 channels (front right, front left, front center, rear right, rear left, and subwoofer) that have been widely used in the film industry.

The AAC is also part of the MPEG-2 audio, i.e., ISO/IEC 13818-7 [11]. It is based on an MDCT filterbank with 1024 bands, and uses Huffman coding for the quantized values. The AAC can also provide multichannel audio (5.1 channels) service.

The coding complexity increases from the MPEG-1 layers I, II, and III, to AC-3 and to AAC. For the same audio quality, more complicated algorithms provide higher compression ratios, or lower bit rates [16]. But they are more vulnerable to transmission errors.

Dolby AC-3 is part of the ATSC DTV standard [13] and has also been adopted by the DVB-T standard [15]. The DVB-T system deployed in Europe implemented MPEG-1 audio layer II to deliver high-quality stereo service. The ISDB-T system adopted AAC as its audio standard.

2.3. Transport Layer

2.3.1. Transport Stream. All TDT systems implemented the MPEG-2 transport system specified in the ISO/IEC 13818-1 standard [11] for data multiplexing. The MPEG-2 transport layer is designed for broadcast and multimedia applications. It has a large packet size and a small amount of overhead to achieve high spectrum efficiency. The MPEG-2 transport packet stream can carry multiplexed compressed video, compressed audio, and data packets.

The MPEG-2 transport stream (TS) packet size is 188 bytes, as shown in Fig. 3. The first byte is the synchronization byte. The next 3 bytes are fixed header bytes for error handling, encryption control, picture identification (PID), priority, and so on. The other 184 bytes are for the adaptation header and payload. The adaptation header is of variable length up to 184 bytes. It is for time synchronization, random-access flagging splicing indication, and so forth.

2.3.2. Program and System Information Protocol and Service Information. The Program and System Information Protocol (PSIP) and service information (SI) are data that are transmitted along with a station's TDT signal to provide TDT receivers important information about the television station and what is being broadcast [17]. In the ATSC system, these data are called PSIP [18]. In the DVB-T and the ISDB-T systems, it is called SI [19,20]. The most important function of PSIP and SI is to provide a method for TDT receivers to identify a TDT station and to determine how a receiver can tune to it. The PSIP and SI also tell the receiver whether multiple program channels are being broadcast and how to find each channel; identify whether the program is closed captioned; and convey program rating information and other data associated with the program. If the TV station inserted the wrong PSIP or SI, the receivers might not be able to find and decode the TDT signal.

4. SERVICES AND COVERAGE

For analog TV coverage planning, the reference receiver setup assumed the use of a directional antenna at a 10-m height at the edge of the coverage area. This means very high antenna gain and directivity (a narrow beamwidth). The 10-m antenna height likely provides a line-of-sight (LoS) path to the transmitter, with resulting high signal strength. This receiver setup is still used for TDT system coverage prediction for fixed-antenna outdoor reception of full data rate transmission. Table 3 [7] provides TDT protection ratios for frequency planning used by different countries and regions.

Since the early 1950s, major population centers have expanded substantially and there have been many highrise buildings and other human-made structures erected throughout urban and suburban areas. In North America, some community bylaws even prohibited the use of outdoor antennas. The widespread use of electrical appliances and high-voltage power lines has substantially increased the level of impulsive noise (especially in the VHF band). All these result in much worse reception conditions, namely, the loss of LoS to the transmitter and the increase of the interference and noise levels. During the analog TV–DTV transition period, the TDT transmission power is further



Figure 3. MPEG-2 transport stream.

	Table 3.	Digital Television	Protection	Ratios (in	dB) for	Frequency	v Planning
--	----------	---------------------------	------------	------------	---------	-----------	------------

System Parameters (Protection Ratios)	Canada	USA	EBU^{a}	Japan ^a
SNR for AWGN channel	+19.5	+15.19	+19.3	+20.1
Cochannel DTV into analog TV	+33.8	+34.44	$+34{\sim}37$	+39
Cochannel analog TV into DTV	+7.2	+1.81	+4	+5
Cochannel DTV into DTV	+19.5	+15.27	+19	+21
Lower adjacent channel DTV into analog TV	-16	-17.43	$-5\sim -11^b$	-6.0
Upper adjacent channel DTV into analog TV	-12	-11.95	$-1 \sim -10^b$	-6.0
Lower adjacent channel analog TV into DTV	-48	-47.33	$-34\sim-37^b$	-31
Upper adjacent channel analog TV into DTV	-49	-48.71	$-38\sim-36^b$	-33
Lower adjacent channel DTV into DTV	-27	-28	-30	-26
Upper adjacent channel DTV into DTV	-27	-26	-30	-27

^aDVB-T (8 MHz, 64 QAM, $R = \frac{2}{3}$). ISDB-T (6 MHz, 64 QAM, $R = \frac{3}{4}$), Analog TV (M/NTSC). ^bDepending on analog TV systems used.

limited to prevent interference into the existing analog TV services. At the same time, DTV has to withstand interference from the analog TV.

For fixed services, TDT has to compete with cable, satellite, MMDS/LMDS, and other communication systems. On the other hand, there is an increasing demand for indoor and mobile television and data services.

4.1. Indoor Reception

Indoor reception presents particularly difficult conditions, because of the lower signal strength due to building penetration losses, which can be as high as 25 dB for the VHF/UHF signals. Indoor signals also suffer from strong static and dynamic multipath distortion, due to reflections from indoor walls, as well as from outdoor structures. The lack of LoS path often results in strong pre-echoes. Nearby traffic and the movement of human bodies or even pets can significantly alter the distribution of indoor signals, causing time-varying echoes and field strength variations.

The indoor signal strength and its distribution are related to many factors, such as building structure (concrete, brick, wood), siding material (aluminum, plastic, wood), insulation material (with or without metal coating), and window material (tinted and metal-coated glasses, multilayer glass).

The indoor settop antenna gain and directivity depend very much on frequency and location. For "rabbit ear" antennas, the measured gain varied from about -10 to -4 dBi. For 5-element logarithmic antennas, the gains are between -15 and +3 dBi. The low height of indoor antennas also results in lower signal strength. Meanwhile, indoor environments sometime experience high levels of impulse noise from power lines and home appliances.

The low receiver antenna height, low antenna gain, and poor building penetration mean that an additional 30-40 dB of signal power is needed for reliable indoor reception compared to outdoor reception. However, the use of the single-frequency networks (SFNs) and receiving antenna diversity can considerably improve the indoor reception. Reducing the data rate and consequently lowering the required SNR can also improve the location availability.

4.2. Mobile Reception

Mobile reception also suffers from low antenna gain and low antenna height. The difference in field strength is larger than 10 dB between a receiving antenna at 1.5 m and one at a height of 10 m. In urban areas, building blockage and shadowing also reduce the signal strength significantly.

Mobile reception requires that the receivers withstand strong and dynamic multipath distortion, Doppler effects, and signal fading. A lower data rate is used, on the order of 50-25% of the rate used for the fixed reception. The system SNR over AWGN is generally under 10 dB.

To guarantee a satisfactory service quality, the transmission system must provide the required field strength within the service area, or a high level of location availability. Single-frequency networks (SFNs) and receiving antenna diversity can definitely improve the service quality of mobile reception.

4.3. Single-Frequency Networks (SFNs) and Diversified Transmission

The SFN or a diversified transmission approach can provide stronger field strength throughout the core coverage area and, therefore, can significantly improve the service availability. The receivers have more than one transmitter from which they can receive the signal (diversity gain). They have better chances of having a strong signal path to a transmitter, which makes it more likely to achieve reliable service.

Optimizing the transmitter network (transmitter density, tower height and location, as well as the transmission power at each transmitter) can result in better coverage with lower total transmit power and better spectrum efficiency. Special measures must be taken to minimize the frequency offset among the repeaters and to flexibly address each transmitter with respect to its exact site, power, antenna height, and the insertion of specific local signal delays.

The key difference between a TDT and an analog TV system is that the TDT can withstand at least 20 dB of DTV–DTV cochannel interference, as shown in Table 3, while the analog TV co-channel threshold of visibility is around 50 dB (30-35 dB for CCIR grade 3). In other words, TDT is 10-30 dB more robust than analog TV, which provides more flexibility for the repeater design and siting.

One drawback of using multiple transmitters is that "active" multipath distortion can occur when coverage from transmitters overlaps. The receiver must be able to deal with these strong echoes. Achieving frequency and time synchronization of multiple transmitters and feeding the same program source to multiple transmitter sites will increase the complexity of the transmission facilities. It could also double the Doppler effects, if the mobile terminal is leaving a reception area from one transmitter and driving toward another transmitter in the overlapping area.

5. CONCLUSION

In comparison to the analog TV services, TDT can provide much better picture and audio qualities. It is robust to multipath distortion and various forms of interference. It is spectrum-efficient and can provide better service quality. It can easily interface with computer systems, and can provide multimedia and data broadcasting services. The TDT will eventually replace all the existing analog TV services.

BIOGRAPHY

Dr. Yiyan Wu received his B.Eng. degree in 1982 from the Beijing University of Posts and Telecommunications, Beijing, China, and M.Eng. and Ph.D. degrees in electrical engineering from Carleton University, Ottawa, Canada, in 1986 and 1990, respectively. He joined Telesat Canada in 1990 as a senior satellite communication systems engineer. Since 1992, he has been a senior research scientist at the Communications Research Centre Canada, where he has been working on digital television and broadband wireless multimedia communication research and standards development. Dr. Wu is a fellow of the IEEE, a member of the editorial board of the proceedings of the IEEE, and an associate editor of the IEEE Transactions on Broadcasting. He has served on many international committees (ATSC, IEEE, ITU) and has been a consultant to many industry and government institutions. He was the recipient of the 1999 IEEE Consumer Electronics Society Chester Sally Paper Award and 2002 Canadian Government Federal Partners in Technology Transfer Innovator Awards for scientific achievement. He is an adjunct professor of Carleton University and the Beijing University of Posts and Telecommunications. Dr. Wu has published more than 150 scientific papers and book chapters.

BIBLIOGRAPHY

- 1. K. B. Benson, ed., *Television Engineering HandBook*, McGraw-Hill, New York, 2000.
- 2. G. Ungerboeck, Trellis coded modulation with redundant signal sets, *IEEE Commun. Mag.* 27: 5–21 (Feb. 1987).
- G. C. Clark and J. B. Cain, Error-Correction Coding for Digital Communications, Plenum Press, New York, 1981.
- J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1995.
- Y. Wu and W. Y. Zou, Orthogonal frequency division multiplexing: A multiple-carrier modulation scheme, *IEEE Trans. Consumer Electron.* 41(3): 392–399 (Aug. 1995).

- S. B. Winstein and P. M. Ebert, Data transmission by frequency division multiplexing using the discrete Fourier transform, *IEEE Trans. Commun.* 19: 628-634 (Oct. 1971).
- 7. Y. Wu et al., Comparison of terrestrial DTV transmission systems: The ATSC 8-VSB, the DVB-T COFDM and the ISDB-T BST-OFDM, *IEEE Trans. Broadcast.* **46**(2): 101–113 (June 2000).
- ATSC, ATSC Digital Television Standard, ATSC Standard A/53, April 1, 2001, http://www.atsc.org/.
- ETS 300 744, Digital Video Broadcasting (DVB); Digital Broadcasting Systems for Television, Sound and Data Services; Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television, ETSI Draft EN 300 744 V1.2.1, (1999-1). http://www.dvb.org/ and http://www.etsi.org/.
- ARIB, Terrestrial Integrated Services Digital Broadcasting (ISDB-T)—Specifications of Channel Coding, Framing Structure, and Modulation, Sept. 28, 1998, http://www.arib.or.jp/arib/english/.
- ISO/IEC 13818, Information Technology Generic Coding of Moving Pictures and Associated Audio Information, Part 1 (Systems), 2 (Video), 3 (Audio), 7 (AAC Audio), 1994.
- ISO/IEC 11172, Information Technology Coding for Moving Pictures and Associated Audio for Digital Storage at about 1.5 Mbits, Part 1 (Systems), 2 (Video), 3 (Audio), 1993.
- ATSC, ATSC Digital Audio Compression Standard (AC/3), ATSC Standard A/52, December 20, 1995.
- 14. ITU-R, Audio Coding for Digital Terrestrial Television Broadcasting, ITU-R Rec. BS.1196, 1995.
- ETSI TR 101 154, Implementation Guidelines for the Use of MPEG-2 Systems, Video and Audio in Satellite, cable and Terrestrial broadcasting Applications, ETSI TR 101 154 V1.4.1 (2000-07).
- G. A. Soulodre et al., Subjective evaluation of state-of-the-art two-channel audio codecs, J. Audio Eng. Soc. 46(3): 164–177 (March 1998).
- A. Allison and K. T. Williams, Channel Branding and Navigation for DTV: Understanding PSIP, NAB, Washington DC, 2000.
- ATSC, Program and System Information Protocol, ATSC Standard A/65A, May 2000.
- ETS 300 468, Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB Systems, ETSI 300 468 ed. 2, Oct. 1996.
- 20. ARIB, Service Information for Digital Broadcasting System, ARIB STD B-10, V3.0, May 2001.

TERRESTRIAL MICROWAVE COMMUNICATIONS

DAVID R. SMITH George Washington University Ashburn, Virgina

1. INTRODUCTION

The basic components required for operating a radio over a microwave link are the transmitter, towers, antennas, and receiver. Transmitter functions typically include multiplexing, encoding, modulation, upconversion from baseband or intermediate frequency (IF) to radiofrequency (RF), power amplification, and filtering for spectrum control. Antennas are placed on a tower or other tall structure at sufficient height to provide a direct, unobstructed lineof-sight (LoS) path between the transmitter and receiver sites. Receiver functions include RF filtering, downconversion from RF to IF amplification at IF equalization, demodulation, decoding, and demultiplexing.

In describing terrestrial microwave, the focus is on digital, line-of-sight, point-to-point microwave systems ($\sim 1-30$ GHz). The design of millimeter-wave (30–300-GHz) radio links is also considered, however. Further, much of the material on line-of-sight propagation, including multipath and interference effects, and link design methodology also applies to the design of mobile and analog radio systems. Many of the design and performance characteristics considered here apply to satellite communications as well.

2. LINE-OF-SIGHT PROPAGATION

The modes of propagation between two radio antennas may include a direct, line-of-sight (LoS) path but also a ground or surface wave that parallels the earth's surface, a sky wave from signal components reflected off the troposphere or ionosphere, a ground reflected path, and a path diffracted from an obstacle in the terrain. The presence and utility of these modes depend on the link geometry, both distance and terrain between the two antennas, and the operating frequency. For frequencies in the microwave band, the LoS propagation mode is the predominant mode available for use; the other modes may cause interference with the stronger LoS path. Line-ofsight links are limited in distance by the curvature of the earth, obstacles along the path, and free-space loss. Average distances for conservatively designed LoS links are 25–30 mi, although distances of ≤ 100 mi have been used. The performance of the LoS path is affected by several phenomena addressed in this section, including free-space loss, terrain, atmosphere, and precipitation. The problem of fading due to multiple paths is addressed in the following section.

2.1. Free-Space Loss

Consider a radio path consisting of isotropic antennas at both transmitter and receiver. An isotropic transmitting antenna radiates its power P_t equally in all directions. In the absence of terrain or atmospheric effects (i.e., free space), the radiated power density is equal at points equidistant from the transmitter. The effective area A of a receiving antenna determines how much of the incident signal from the transmitting antenna is collected by the receiving antenna is given by

$$A = \frac{\lambda^2 G_r}{4\pi} \tag{1}$$

where λ = wavelength in meters and G_r is the ratio of the receiving antenna power gain to the corresponding gain of

an isotropic radiator. The power received by the receiving antenna at a distance r in watts, P_r , is then given by

$$P_r = \frac{P_t G_t}{4\pi r^2} \cdot \frac{\lambda^2 G_r}{4\pi} \tag{2}$$

If we assume that both the transmitting and receiving antennas are isotropic radiators, then $G_t = G_r = 1$, and the ratio of received to transmitted power becomes the free-space path loss

$$l_{\rm fs} = \frac{P_r}{P_t} = \frac{\lambda^2}{(4\pi r)^2} \tag{3}$$

The free-space path loss is then determined as the ratio of the received power to the transmitted power and is given in decibels by

$$\begin{split} L_{\rm fs} &= 10 \log_{10}(l_{\rm fs}) = 10 \log_{10}\left(\frac{P_r}{P_t}\right) \\ &= 32.44 + 20 \log_{10}(D_{\rm km}) + 20 \log_{10}(f_{\rm MHz}) \ {\rm dB} \quad (4) \end{split}$$

where $r = D_{\rm km}$ is the distance in kilometers and $f_{\rm MHz} = 300/\lambda_{\rm meters}$ is the signal frequency in megahertz (MHz). Note that the doubling of either frequency or distance causes a 6-dB increase in path loss.

2.2. Terrain Effects

Obstacles along a LoS radio path can cause the propagated signal to be reflected or diffracted, resulting in path losses that deviate from the free space value. This effect stems from electromagnetic wave theory, which postulates that a wavefront diverges as it advances through space. A radio beam that just grazes the obstacle is diffracted, with a resulting obstruction loss whose magnitude depends on the type of surface over which the diffraction occurs. A smooth surface, such as water or flat terrain, produces the maximum obstruction loss at grazing. A sharp projection, such as a mountain peak or even trees, produces a knifeedge effect with minimum obstruction loss at grazing. Most obstacles in the radio path produce an obstruction loss somewhere between the limits of smooth earth and knife edge.

2.2.1. Reflections. When the obstacle is below the optical LoS path, the radio beam can be reflected to create a second signal at the receiving antenna. Reflected signals can be particularly strong when the reflection surface is smooth terrain or water. Since the reflected signal travels a longer path than does the direct signal, the reflected signal may arrive out of phase with the direct signal. The degree of interference at the receiving antenna from the reflected signal depends on the relative signal levels and phases of the direct and reflected signals.

At the point of reflection, the indirect signal undergoes attenuation and phase shift, which is described by the reflection coefficient R, where

$$R = \rho \exp(-j\phi) \tag{5}$$

The magnitude ρ represents the change in amplitude, and ϕ is the phase shift on reflection. The values of ρ and ϕ depend on the wave polarization (horizontal or vertical), angle of incidence ψ dielectric constant of the reflection surface, and wavelength λ of the radio signal. The mathematical relationship has been developed elsewhere [1] and will not be covered here. For microwave frequencies, however, two general cases should be mentioned:

- 1. For horizontally polarized waves with small angle of incidence, R = -1 for all terrain, such that the reflected signal suffers no change in amplitude but has a phase change of 180° .
- 2. If the polarization is vertical with grazing incidence, R = -1 for all terrain. With increasing angle of incidence, the reflection coefficient magnitude decreases, reaching zero in the vicinity of $\psi = 10^{\circ}$.

To examine the problem of interference from reflection, we first simplify the analysis by neglecting the effects of the curvature of the earth's surface. Then, when the reflection surface is flat earth, the geometry is as illustrated in Fig. 1a, with transmitter (Tx) at height h_1 and receiver (Rx) at height h_2 separated by a distance D and the angle of reflection equal to the angle of incidence ψ . Using plane geometry and algebra, the path difference δ between the reflected and direct signals can be given by

$$\delta = (r_2 + r_1) - r = \frac{2h_1h_2}{D}$$
(6)

The overall phase change experienced by the reflected signal relative to the direct signal is the sum of the phase difference due to the pathlength difference δ and the phase ϕ due to the reflection. The total phase shift is therefore

$$\gamma = \frac{2\pi}{\lambda} \frac{2h_1 h_2}{D} + \phi \tag{7}$$

At the receiver, the direct and reflected signals combine to form a composite signal with field strength E_C . By the



Figure 1. Geometry of two-path propagation: (a) flat-earth geometry; (b) two-ray geometry.

simple geometry of Fig. 1b and using the law of cosines, we obtain

$$E_C = E_D^2 + E_R^2 + 2E_D E_R \cos\gamma \tag{8}$$

where E_D = field strength of direct signal

- E_R = field strength of reflected signal
 - $\rho = \text{magnitude of reflection coefficient} = E_R/E_D$
 - γ = phase difference between direct and reflected
 - signal as given by Eg. (7)

The composite signal is at a minimum, from (8), when $\gamma = (2n + 1)\pi$, where *n* is an integer. Similarly, the composite signal is at a maximum when $\gamma = 2n\pi$. As noted earlier, the phase shift ϕ due to reflection is usually around 180° for microwave paths since the angle of incidence on the reflection surface is typically quite small. For this case, the received signal minima, or nulls, occur when the path difference is an even multiple of a half wavelength, or

$$\delta = 2n\left(\frac{\lambda}{2}\right)$$
 for minima (9)

The maxima, or peaks, for this case occur when the path difference is an odd multiple of a half-wavelength, or

$$\delta = \frac{(2n+1)\lambda}{2} \quad \text{for maxima} \tag{10}$$

2.2.2. Fresnel Zones. The effects of reflection and diffraction on radiowaves can be more easily seen by using the model developed by Fresnel for optics. Fresnel accounted for the diffraction of light by postulating that the cross section of an optical wavefront is divided into zones of concentric circles separated by half-wavelengths. These zones alternate between constructive and destructive interference, resulting in a sequence of dark and light bands when diffracted light is viewed on a screen. When viewed in three dimensions, as necessary for determining path clearances in LoS radio systems, the Fresnel zones become concentric ellipsoids. The first Fresnel zone is that locus of points for which the sum of the distances between the transmitter and receiver and a point on the ellipsoid is exactly one half-wavelength longer than the direct path between the transmitter and receiver. The nth Fresnel zone consists of that set of points for which the difference is n half-wavelengths. The radius of the *n*th Fresnel zone at a given distance along the path is given by

$$F_n = 17.3 \left(\frac{nd_1d_2}{fD}\right)^{1/2} \text{ meters}$$
(11)

where d_1 = distance from transmitter to a given point along the path (km)

- $d_2 = \text{distance from receiver to the same point}$ along the path (km)
- f =frequency (GHz)
- $D = \text{pathlength} (\text{km}) (D = d_1 + d_2)$

As an example, Fig. 2 shows the first three Fresnel zones for an LoS path of length (D) 40 km and frequency (f) 8 GHz. The distance h represents the clearance between the LoS path and the highest obstacle along the terrain.



Figure 2. Fresnel zones: (a) Fresnel zones for an 8-GHz, 40-km LoS path; (b) attenuation versus path clearance.

Using Fresnel diffraction theory, we can calculate the effects of path clearance on transmission loss (see Fig. 2b). The three cases shown in Fig. 2b correspond to different reflection coefficient values as determined by differences in terrain roughness. The curve marked R = 0 represents the case of knife-edge diffraction, where the loss at a grazing angle (zero clearance) is equal to 6 dB. The curve marked R = -1.0 illustrates diffraction from a smooth surface, which produces a maximum loss equal to 20 dB at grazing. In practice, most microwave paths have been found to have a reflection coefficient magnitude of 0.2-0.4; thus the curve marked R = -0.3 represents the ordinary path [2]. For most paths, the signal attenuation becomes small with a clearance of 0.6 times the first Fresnel zone radius. Thus microwave paths are typically sited with a clearance of at least $0.6F_1$.

The fluctuation in signal attenuation observed in Fig. 2b is due to alternating constructive and destructive interference with increasing clearance. Clearance at odd-numbered Fresnel zones produces constructive interference since the delayed signal is in phase with the direct signal; with a reflection coefficient of -1.0, the direct and

delayed signals sum to a value 6 dB higher than free-space loss. Clearance at even-numbered Fresnel zones produces destructive interference since the delayed signal is out of phase with the direct signal by a multiple of $\lambda/2$; for a reflection coefficient of -1.0, the two signals cancel each other. As indicated in Fig. 2b, the separation between adjacent peaks or nulls decreases with increasing clearance, but the difference in signal strength decreases with increasing Fresnel zone numbers.

2.3. Atmospheric Effects

Radiowaves travel in straight lines in free space, but they are bent, or refracted, when traveling through the atmosphere. Bending of radiowaves is caused by changes with altitude in the index of refraction, defined as the ratio of propagation velocity in free space to that in the medium of interest. Normally the refractive index decreases with altitude, meaning that the velocity of propagation increases with altitude, causing radiowaves to bend downward. In this case, the radio horizon is extended beyond the optical horizon. The index of refraction n varies from a value of 1.0 for free space to approximately 1.0003 at the surface of the earth. Since this refractive index varies over such a small range, it is more convenient to use a scaled unit, N, which is called *radio refractivity* and defined as

$$N = (n-1)10^6 \tag{12}$$

Thus N indicates the excess over unity of the refractive index, expressed in millionths. When n = 1.0003, for example, N has a value of 300. Owing to the rapid decrease of pressure and humidity with altitude and the slow decrease of temperature with altitude, N normally decreases with altitude and tends to zero.

To account for atmospheric refraction in path clearance calculations, it is convenient to replace the true earth radius a by an effective earth radius a_e and to replace the actual atmosphere with a uniform atmosphere in which radiowaves travel in straight lines. The ratio of effective to true earth radius is known as the k factor:

$$k = \frac{a_e}{a} \tag{13}$$

By application of Snell's law in spherical geometry, it may be shown that as long as the change in refractive index is linear with altitude, the k factor is given by

$$k = \frac{1}{1 + a(dn/dh)} \tag{14}$$

where dn/dh is the rate of change of refractive index with height. It is usually more convenient to consider the gradient of N instead of the gradient of n. Making the substitution of dN/dh for dn/dh and also entering the value of 6370 km for a into (14) yields the following:

$$k = \frac{157}{157 + (dN/dh)} \tag{15}$$

where dN/dh is the *N* gradient per kilometer. Under most atmospheric conditions, the gradient of *N* is negative and constant and has a value of approximately

$$\frac{dN}{dh} = -40 \text{ units/km}$$
(16)

Substituting (14) into (13) yields a value of $k = \frac{4}{3}$, which is commonly used in propagation analysis. An index of refraction that decreases uniformly with altitude resulting in $k = \frac{4}{3}$ is referred to as *standard refraction*.

2.3.1. Anomalous Propagation. Weather conditions may lead to a refractive index variation with height that differs significantly from the average value. In fact, atmospheric refraction and corresponding k factors may be negative, zero, or positive. The various forms of refraction are illustrated in Fig. 3 by presenting radio paths over both true earth and effective earth. Note that radiowaves become straight lines when drawn over the effective earth radius. Standard refraction is the average condition observed and results from a well-mixed atmosphere. The other refractive conditions illustrated



Figure 3. Various forms of atmospheric refraction; (a) standard refraction $(k = \frac{4}{3})$; (b) subrefraction (0 < k < 1); (c) superrefraction $(2 < k < \infty)$; (d) ducting (k < 0).

in Fig. 3—including subrefraction, superrefraction, and ducting—are observed a small percentage of the time and are collectively referred to as *anomalous propagation*.

Subrefraction (k < 1) leads to the phenomenon known as *inverse bending* or "earth bulge", illustrated in Fig. 3b. This condition arises because of an increase in refractive index with altitude and results in an upward bending of radiowaves. Substandard atmospheric refraction may occur with the formation of fog, as cold air passes over a warm earth, or with atmospheric stratification, as occurs at night. The effect produced is likened to the bulging of the earth into the microwave path that reduces the path clearance or obstructs the LoS path.

Superrefraction (k > 2) causes radiowaves to refract downward with a curvature greater than normal. The result is an increased flattening of the effective earth. For the case illustrated in Fig. 3c the effective earth radius is infinity-that is, the earth reduces to a plane. From Eq. (15) it can be seen that an N gradient of -157 units per kilometer yields a k equal to infinity. Under these conditions radiowaves are propagated at a fixed height above the earth's surface, creating unusually long propagation distances and the potential for overreach interference with other signals occupying the same frequency allocation. Superrefractive conditions arise when the index of refraction decreases more rapidly than normal with increasing altitude, which is produced by a rise in temperature with altitude, a decrease in humidity, or both. An increase in temperature with altitude, called a temperature inversion, occurs when the temperature of

the earth's surface is significantly less than that of the air, which is most commonly caused by cooling of the earth's surface through radiation on clear nights or by movement of warm dry air over a cooler body of water.

A more rapid decrease in refractive index gives rise to more pronounced bending of radiowaves, in which the radius of curvature of the radiowave is smaller than the earth's radius. As indicated in Fig. 3d, the rays are bent to the earth's surface and then reflected upward from it. With multiple reflections, the radiowaves can cover large ranges far beyond the normal horizon. In order for the radiowave's bending radius to be smaller than the earth's radius, the N gradient must be less than -157 units per kilometer. Then, according to (15), the k factor and effective earth radius both become negative quantities. As illustrated in Fig. 3d, the effective earth is approximated by a concave surface. This form of anomalous propagation is called ducting because the radio signal appears to be propagated through a waveguide, or duct. A duct may be located along or it elevated above the earth's surface. The meteorological conditions responsible for either surface or elevated ducts are similar to conditions causing superrefractivity. With ducting, however, a transition region between two differing air masses creates a trapping layer. In ducting conditions, refractivity N decreases with increasing height in an approximately linear fashion above and below the transition region, where the gradient departs from the average. In this transition region, the gradient of Nbecomes steep.

2.3.2. Atmospheric Absorption. For frequencies above 10 GHz, attenuation due to atmospheric absorption becomes an important factor in radio-link design. The two major atmospheric gases contributing to attenuation are water vapor and oxygen. Studies have shown that absorption peaks occur in the vicinity of 22.3 and 187 GHz due to water vapor and in the vicinity of 60 and 120 GHz for oxygen [3]. The calculation of specific attenuation produced by either oxygen or water vapor is complex, requiring computer evaluation for each value of temperature, pressure, and humidity. Formulas that approximate specific attenuation may be found in ITU-R Rep. 721-2 [4]. At millimeter wavelengths (30–300 GHz), atmospheric absorption becomes a significant problem. To obtain maximum propagation range, frequencies around the absorption peaks are to be avoided. On the other hand, certain frequency bands have relatively low attenuation. In the millimeter-wave range, the first two such bands, or windows, are centered at approximately 36 and 85 GHz.

2.3.3. Rain Attenuation. Attenuation due to rain and suspended water droplets (fog) can be a major cause of signal loss, particularly for frequencies above 10 GHz. Rain and fog cause a scattering of radiowaves that results in attenuation. Moreover, for the case of millimeter wavelengths where the raindrop size is comparable to the wavelength, absorption occurs and increases attenuation. The degree of attenuation on an LoS link is a function of (1) the point rainfall rate distribution; (2) the specific attenuation, which relates rainfall rates to point attenuations; and (3) the effective pathlength,

which is multiplied by the specific attenuation to account for the length of the path. Heavy rain, as found in thunderstorms, produces significant attenuation, particularly for frequencies above 10 GHz. The point rainfall rate distribution gives the percentage of a year that the rainfall rate exceeds a specified value. Rainfall rate distributions depend on climatologic conditions and vary from one location to another. To relate rainfall rates to a particular path, measurements must be made by use of rain gauges placed along the propagation path. In the absence of specific rainfall data along a specific path, it becomes necessary to use maps of rain climate regions, such as those provided by ITU-R Rep. 563.3 [4].

To counter the effects of rain attenuation, it should first be noted that neither space nor frequency diversity is effective as protection against rainfall effects. Measures that are effective, however, include increasing the fade margin, shortening the pathlength, and using a lower-frequency band. A mathematical model for rain attenuation is found in ITU-R Rep. 721-3 [4] and summarized here. The specific attenuation is the loss per unit distance that would be observed at a given rain rate, or

$$\gamma_R = kR^\beta \tag{17}$$

where γ_R is the specific attenuation in dB/km and R is the point rainfall rate in mm/h. The values of k and β depend on the frequency and polarization, and may be determined from tabulated values in ITU-R Rep. 721-2. Because of the nonuniformity of rainfall rates within the cell of a storm, the attenuation on a path is not proportional to the pathlength; instead it is determined from an effective pathlength given in [5] as

$$L_{\rm eff} = \frac{L}{1 + (R - 6.2)L/2636}$$
(18)

Now, to determine the attenuation for a particular probability of outage, the ITU-R method calculates the attenuation $A_{0.01}$ that occurs 0.01% of a year and uses a scaling law to calculate the attenuation A, at other probabilities

$$A_{0.01} = \gamma_R L_{\text{eff}} \tag{19}$$

where γ_R and $L_{\rm eff}$ are determined for the 0.01% rainfall rate. The value of the 0.01% rainfall rate is obtained from world contour maps of 0.01% rainfall rates found in ITU-R Rep. 563-4.

2.4. Path Profiles

In order to determine tower heights for suitable path clearance, a profile of the path must be plotted. The path profile is obtained from topographic maps that should have a scale of 1:50,000 or less. For LoS links under 70 km in length, a straight line may be drawn connecting the two endpoints. For longer links, the great circle path must be calculated and plotted on the map. The elevation contours are then read from the map and plotted on suitable graph paper, taking special note of any obstacles along the path. The path profiling process may be fully automated by use of CD-ROM technology and an appropriate computer program. CD-ROM data storage disks are available that contain a global terrain elevation database from which profile points can be automatically retrieved, given the latitudes and longitudes of the link endpoints [12].

The path profile may be plotted on special graph paper that depicts the earth as curved and the transmitted ray as a straight line or on rectilinear graph paper that depicts the earth as flat and the transmitted ray as a curved line. The use of linear paper is preferred because it eliminates the need for special graph paper, permits the plotting of rays for different effective earth radius, and simplifies the plotting of the profile. Figure 4 is an example of a profile plotted on linear paper for a 40-km, 8-GHz radio link.

The use of rectilinear paper, as suggested, requires the calculation of the earth bulge at a number of points along the path, especially at obstacles. This calculation then accounts for the added elevation to obstacles due to curvature of the earth. Earth bulge in meters may be calculated as

$$h = \frac{d_1 d_2}{12.76} \tag{20}$$

- where d_1 = distance from one end of path to point being calculated (km)
 - $d_2 = {
 m distance from same point to other end of the path (km)}$

As indicated earlier, atmospheric refraction causes ray bending, which can be expressed as an effective change in earth radius by using the k factor. The effect of refraction on earth bulge can be handled by adding the k factor to the denominator in (20) or, for d_1 and d_2 in miles and hin feet

$$h = \frac{d_1 d_2}{1.5 k}$$
(21)

To facilitate path profiling, Eq. (21) may he used to plot a curved ray template for a particular value of k and for use with a flat-earth profile. Alternatively, the earth bulge can be calculated and plotted at selected points that represent the clearance required below a straight line drawn between antennas; when connected together, these points form a smooth parabola whose curvature is determined by the choice of k.



Figure 4. Example of a LoS path profile plotted on linear paper.

In path profiling, the choice of k factor is influenced by its minimum value expected over the path and the path availability requirement. With lower values of k, earth bulging becomes pronounced and antenna height must be increased to provide clearance. To determine the clearance requirements, the distribution of k values is required; it can be found by meteorological measurements [4]. This distribution of k values can be related to path availability by selecting a k whose value is exceeded for a percentage of the time equal to the availability requirement.

Apart from the k factor, the Fresnel zone clearance must be added. Desired clearance of any obstacle is expressed as a fraction, typically 0.3 or 0.6, of the first Fresnel zone radius. This additional clearance is then plotted on the path profile, shown as a small tickmark on Fig. 4, for each point being profiled. Finally, clearance should be provided for trees (nominally 15 m) and additional tree growth (nominally 3 m) or, in the absence of trees, for smaller vegetation (nominally 3 m).

The clearance criteria can thus be expressed by specific choices of k and fraction of first Fresnel zone. Here is one set of clearance criteria that is commonly used for highly reliable paths [7]:

- 1. Full first Fresnel zone clearance for $k = \frac{4}{3}$
- 2. 0.3 first Fresnel zone clearance for $k = \frac{2}{3}$

whichever is greater. Over the majority of paths, the clearance requirements of criterion 2 will be controlling. Even so, the clearance should be evaluated by using both criteria along the entire path.

Path profiles are easily obtained using digital terrain data and standard computer software. The United States Geological Survey (USGS) database called *Digital Elevation Models* (DEM) contains digitized elevation data versus latitude and longitude throughout the United States. A similar database for the world has been developed by the National Imagery and Mapping Agency (NIMA) called *Digital Terrain Elevation Data* (DTED). The USGS DEM data and DEM viewer software are available from the USGS web site.

The smallest spacing between elevation points in the USGS is 100 ft (~30 m), while larger spacings of 3 arcseconds (300 ft north-south and 230 east-west) are also available. The USGS 3-arcsecond data are provided in $1^{\circ}\times1^{\circ}$ blocks for the United States. The 1° DEM is also referred to as DEM 250 because these data were collected from 1:250,000-scale maps. DTED level 1 data are identical to DEM 250 except for format differences. Even at the 100-ft spacing, there is a significant number and diversity of terrain features excluded from the database that could significantly impact a propagating wave at microwave radiofrequencies. These terrain databases also contain some information regarding land use/land clutter (LULC). The LULC data indicate geographic areas covered by foliage, buildings, and other surface details. The USGS data contain LULC information that could be used to estimate foliage losses in rural areas and manmade noise levels near built-up areas. In addition, they also contain other terrain features such as roadways, bodies of water, and other information directly relevant to the wireless environment.

3. MULTIPATH FADING

Fading is defined as variation of received signal level with time due to changes in atmospheric conditions. The propagation mechanisms that cause fading include refraction, reflection, and diffraction associated with both the atmosphere and terrain along the path. The two general types of fading, referred to as *multipath* and *power fading*, are illustrated by the recordings of RF received signal levels shown in Fig. 5.

Power fading, sometimes called attenuation fading, results mainly from anomalous propagation conditions, such as (see Fig. 3) subrefraction k < 1), which causes blockage of the path due to the effective increase in earth bulge; superrefraction (k > 2), which causes pronounced ray bending and decoupling of the signal from the receiving antenna; and ducting (k < 0), in which the radio beam is trapped by atmospheric layering and directed away from the receiving antenna. Rainfall also contributes to power fading, particularly for frequencies above 10 GHz. Power fading is characterized as slowly varying in time, usually independent of frequency, and causing long periods of outages. Remedies include greater antenna heights for subrefractive conditions, antenna realignment for superrefractive conditions, and added link margin for rainfall attenuation.

Multipath fading arises from destructive interference between the direct ray and one or more reflected or refracted rays. These multiple paths are of different lengths and have varied phase angles on arrival at the receiving antenna. These various components sum to produce a rapidly varying, frequency-selective form of fading. Deep fades occur when the primary and secondary rays are equal in amplitude but opposite in phase, resulting in signal cancellation and a deep amplitude null. Between deep fades, small amplitude fluctuations are observed that are known as *scintillation*; these fluctuations are due to weak secondary rays interfering with a strong direct ray.

Multipath fading is observed during periods of atmospheric stratification, where layers exist with different refractive gradients. The most common meteorological cause of layering is a temperature inversion, which commonly occurs in hot, humid, still, windless conditions, especially in late evening, at night, and in early morning. Since these conditions arise during the summer, multipath fading is worst during the summer season. Multipath fading can also be caused by reflections from flat terrain or a body of water. Hence multipath fading conditions are most likely to occur during periods of stable atmosphere and for highly reflective paths. Multipath fading is thus a function of pathlength, frequency, climate, and terrain. Techniques used to deal with multipath fading include the use of diversity, increased fade margin, and adaptive equalization.

3.1. Statistical Properties of Fading

The random nature of multipath fading suggests a statistical approach to its characterization. The statistical parameters commonly used in describing fading are

- Probability (or percentage of time) that the LoS link is experiencing a fade below threshold
- Average fade duration and probability of fade duration greater than a given time
- Expected number of fades per unit time

The terms to be used are defined in graph form in Fig. 6. The threshold L is the signal level corresponding to the minimum acceptable signal-to-noise ratio or, for digital transmission, the maximum acceptable probability of error. The difference between the normal received signal level and threshold is the fade margin. A *fade* is defined as the downward crossing of the received signal through



Figure 5. Example of multipath and power fading for LoS link.



the threshold. The time spent below threshold for a given fade is then the fade duration.

For LoS links, the probability distribution of fading signals is known to be related to and limited by the Rayleigh distribution, which is well known and is found by integrating the curve shown in Fig. 7. The Rayleigh probability density function is given by

$$p(r) = \begin{cases} (r/\sigma^2)e^{-r^2/2\sigma^2} & 0 \le r < \infty\\ 0 & \text{otherwise} \end{cases}$$
(22)

for envelope amplitude r and mean square amplitude σ^2 . The Rayleigh distribution function has the form

$$P(r_0) = \Pr(r \le r_0) = \int_0^{r_0} p(r) \, dr = 1 - \exp(-\frac{r_0^2}{2\sigma^2})$$
(23)

For relative envelope amplitude $R = r/\sigma \sqrt{2}$ and relative threshold amplitude $L = r_0/\sigma \sqrt{2}$, the distribution function becomes

$$P(R < L) = 1 - \exp(-L^2)$$
(24)

An approximation to (24) valid for small values of L (representing deep fades) is

$$P(R < L) \approx L^2 \quad \text{for} \quad L < 0.1 \tag{25}$$



Figure 7. Rayleigh and Ricean distributions.

Figure 6. Definition of fading terms.

Fading probabilities are more conveniently expressed in terms of the fade margin F in decibels by letting $F = -20 \log L$. Then

$$P(R < L) = 10^{-F/10} \tag{26}$$

Actual observations of multipath fading indicate that in the region of deep fades, amplitude distributions have the same slope as the Rayleigh distribution but displaced. This characteristic corresponds to the special case of the Ricean distribution where a direct (or specular) component exists that is equal to or less than the Rayleigh fading component [8]. Thus, for deep fading, the distribution function becomes

$$P(R < L) = d(1 - \exp(-L^2))$$
(27)

The parameter d that modifies the Rayleigh distribution has been termed a multipath occurrence factor. Experimental results of Barnett [9] show that

$$d = \frac{abD^3f}{4} \times 10^{-5}$$
 (28)

where D = pathlength(mi)

- f =frequency (GHz)
- a =terrain factor: = 4 for overwater or flat terrain;
 - = 1 for average terrain;
 - $=\frac{1}{4}$ for mountainous terrain
- $b = \text{climate factor:} = \frac{1}{2}$ for hot, humid climate;
 - $=\frac{1}{4}$ for average; temperate climate; $=\frac{1}{8}$ for cool, dry climate

Combining this factor with the basic Rayleigh probability of (14) results in the following overall expression for probability of outage due to fading deeper than the fade margin:

$$P(o) = d10^{-F/10} = \frac{abD^3 f \times 10^{-5}}{4} (10^{-F/10})$$
(29)

The Rayleigh distribution given by (14) is the limiting value for multipath fading. Note that the distributions all have a slope of 10 dB per decade of probability.

3.2. Diversity Improvement

Diversity is used in LoS radio links to protect against either equipment failure or multipath fading. Here we

2564 TERRESTRIAL MICROWAVE COMMUNICATIONS

consider the improvement in multipath fading afforded by the two most commonly used diversity techniques:

- *Space diversity*, which provides two signal paths by use of vertically separated receiving antennas
- *Frequency diversity*, which provides two signal frequencies by use of separate transmitter-receiver pairs

The degree of improvement provided by diversity depends on the degree of correlation between the two fading signals. In practice, because of limitations in allowable antenna separation or frequency spacing, the fading correlation tends to be high. Fortunately, improvement in link availability remains quite significant even for high correlation. To derive the diversity improvement, we begin with the joint Rayleigh probability distribution function to describe the fading correlation between diversity signals, given by

$$P(R_1 < L, R_2 < L) = \frac{L^4}{1 - k^2} \quad \text{(for small } L\text{)} \tag{30}$$

where R_1 and R_2 are signal levels for diversity channels 1 and 2 and k^2 is the correlation coefficient. By experimental results, empirical expressions for k^2 have been established that are a function of antenna separation or frequency spacing, wavelength, and pathlength.

3.2.1. Space Diversity Improvement Factor. Vigants [10] has developed the following expression for k^2 in space diversity links:

$$k^2 = 1 - \frac{S^2}{2.75D\lambda}$$
(31)

where S = antenna separation; D = pathlength; λ = wavelength; and S, D, and λ are in the same units. A more convenient expression is the space diversity improvement factor, given by

$$I_{\rm SD} = \frac{P(R_1 < L)}{P(R_1 < L, R_2 < L)} \cong \frac{L^2}{L^4/(1 - k^2)} = \frac{1 - k^2}{L^2} \quad (32)$$

Using Vigants' expression for k^2 [Eq. (31)], we obtain

$$I_{\rm SD} = \frac{S^2}{2.75 \, D\lambda L^2} \tag{33}$$

When *D* is given in miles, *S* in feet, and λ in terms of carrier frequency *f* in gigahertz, *I*_{SD} can be expressed as

$$I_{\rm SD} = \frac{(7.0 \times 10^{-5})fS^2}{D} (10^{F/10}) \tag{34}$$

where F is the fade margin associated with the second antenna.

3.2.2. Frequency Diversity Improvement Factor. Using experimental data and a mathematical model, Vigants and Pursley [11] have developed an improvement factor for frequency diversity, given by

$$I_{\rm FD} = \frac{(50\Delta f)}{fD} (10^{F/10})$$
(35)

where f = frequency (GHz) $\Delta f =$ frequency separation (GHz) F = fade margin (dB) D = pathlength (mi)

3.2.3. Effect of Diversity on Fading Statistics. The effect of diversity improvement, I_d , on probability of outage due to fading can be expressed as

$$P(o) = P_d \frac{(o)}{I_d} \tag{36}$$

where $P_d(o)$ is the outage probability of simultaneous fading in the two diversity signals. For space diversity, substituting Eqs. (29) and (34) into (36), we obtain

$$P_{\rm sd}(o) = \frac{abD^4}{28S^2} 10^{-F/5} \tag{37}$$

where the fade margins on the two antennas are assumed equal. Likewise, for frequency diversity we obtain

$$P_{\rm fd}(o) = (5 \times 10^{-8}) \, \frac{abf^3 D^4}{\Delta f} \, 10^{-{\rm F}/5} \tag{38}$$

3.3. Frequency-Selective Fading

The first experiences with wideband digital radios revealed that measured error performance fell far short of the performance predicted by the fiat fading model assumed in our discussions so far. This result is due to the presence of frequency-selective fading during which the amplitude and group delay characteristics become distorted. For digital signals, this distortion leads to intersymbol interference that, in turn, degrades the system error rate. This degradation is directly proportional to system bit rate, since higher bit rates mean smaller pulsewidths and greater susceptibility to intersymbol interference. Previously, in analog radio transmission, frequency-selective fading caused intermodulation distortion, but this effect was always secondary when compared to the received signal power. For digital radio systems, however, the traditional fade depth is found to be a poor indicator of error rate.

When the amplitude of the received signal is plotted versus frequency, as shown in Fig. 8, deep amplitude notches appear when the direct ray is out of phase with the indirect rays. These notches are separated in frequency by $1/\tau$, where τ is the time delay between the direct and indirect rays. The notch depth is determined by the relative amplitude of the direct and indirect rays. When an amplitude notch or slope appears in the band of a radio channel, degradation in error rate can be expected. This variation of amplitude with frequency, known as *amplitude dispersion*, is often the main source of degradation in digital radio systems.

3.3.1. Channel Models. Both low-order power series [12] and multipath transfer functions [13] have been used to model the effects of frequency-selective fading. Several multipath transfer function models have been developed, usually based on the presence of two [14] or three [13]





Figure 8. Multipath fading effect.

rays. In general, the multipath channel transfer function can be written as

$$H(\omega) = 1 + \sum_{i=1}^{n} \beta_i \exp(j\omega\tau_i)$$
(39)

where the direct ray has been normalized to unity and the β_i and τ_i are amplitude and delay of the interfering rays relative to the direct ray. The two-ray model can thus be characterized by two parameters, β and τ . In this case, the amplitude of the resultant signal is

$$R = (1 + \beta^2 + 2\beta \cos \omega \tau)^{l/2} \tag{40}$$

and the phase of the resultant is

$$\phi = \arctan \frac{\beta \sin \omega \tau}{1 + \beta \cos \omega \tau} \tag{41}$$

Although the two-ray model is easy to understand and apply, most multipath propagation research points toward the presence of three (or more) rays during fading conditions. Out of this research, Rummler's three-ray model [13] is the most widely accepted.

3.3.2. Dispersive Fade Margin. The effects of dispersion due to frequency-selective fading are conveniently characterized by the dispersive fade margin (DFM), defined as in Fig. 6 but where the source of degradation is distortion rather than additive noise. The DFM can be estimated from M-curve signatures. This signature method of evaluation has been developed to determine radio sensitivity to multipath dispersion using analysis [15], computer simulation, or laboratory simulation [16]. The signature approach is based on an assumed fading model (e.g., two-ray or three-ray). The parameters of the fading model are varied over their expected ranges, and the radio performance is analyzed or recorded for each setting. To develop the signature, the parameters are adjusted to provide



a threshold error rate (say, 1×10^{-7}) and a plot of the parameter settings is made to delineate the outage area. A typical signature developed by using the two-ray model is shown in Fig. 9. The area under the signature curve corresponds to conditions for which the bit error rate exceeds the threshold error rate; the area outside the signature corresponds to a BER that is less than the threshold value. The *M* shape of the curve (hence the term *M* curve) indicates that the radio is less susceptible to fades in the center of the band than to off-center fades. As the notch moves toward the band edges, greater notch depth is required to produce the threshold error rate. If the notch lies well outside the radioband, no amount of fading will cause an outage.

3.3.3. Improvements Due to Diversity and Equalization. Both diversity and adaptive equalization can be used, separately or together, to improve digital radio performance in the presence of frequency-selective fading. Diversity reduces the probability of in-band dispersion. Adaptive equalization reduces the in-band difference between the minimum- and maximum-amplitude values and, depending on the type of equalizer, reduces the inband difference between group delay values also. In many instances, both diversity and equalization have been necessary to meet performance objectives. Interestingly, the combined improvement obtained by simultaneous use of diversity and equalization has been found to be larger than the product of the individual improvements. This synergistic effect has been reported in several experiments [18,19], where the added improvement has resulted from the diversity combiner's ability to replace in-band notches with slopes that are easier to equalize. Field measurements of EER distributions for a 6-GHz, 90-Mbps, 8-PSK radio on a 37.3-mi link are shown in Fig. 10 [20]. This system was tested in four configurations: unprotected, with adaptive equalization, with space diversity, and with equalization plus diversity. These results indicate a synergistic effect, with large improvement observed in the combination of equalization and diversity.



Figure 10. BER Distributions for 6-GHz, 90-Mbps, 8-PSK, 37.3-mi link showing effects of equalization and diversity [20].

4. FREQUENCY ALLOCATIONS AND INTERFERENCE EFFECTS

The design of a radio system must include a frequency allocation plan, which is subject to approval by the local frequency regulatory authority. In the United States, radio channel assignments are controlled by the Federal Communications Commission (FCC) for commercial carriers and by the National Telecommunications and Information Administration (NTIA) for government systems. Figure 11 shows the spectrum allocation process in the United States. The FCC's regulations for use of microwave spectrum establish eligibility rules, permissible-use rules, and technical specifications. There are four principal users who either share or exclusively use a particular spectrum allocation: common carriers, broadcasters, cable TV operators, and private companies. FCC regulatory specifications are intended to protect against interference and to promote spectral efficiency. Equipment-type acceptance regulations include transmitter power limits, frequency stability, out-of-channel emission limits, and antenna directivity. For digital microwave, specifications are added



Figure 11. Spectrum allocation process in the United States.

for digital modulation, spectral efficiency in bps/Hz, and voice channel capacity [21].

The International Telecommunications Union Radio Committee (ITU-R) issues recommendations on radio channel assignments for use by national frequency allocation agencies. Although the ITU-R itself has no regulatory power, it is important to realize that ITU-R recommendations are usually adopted on a worldwide basis. With regard to digital microwave systems, the ITU-R has issued recommendations beginning with the 1982 Plenary Assembly.

The usual practice in frequency channel assignments for a particular frequency band is to separate transmit (GO) and receive (RETURN) frequencies by placing all GO channels in one half of the band and all RETURN channels in the other half. With this approach, all transmitters on a given station are in either the upper or lower half of the band, with receivers in the remaining half. Within each half-band, adjacent channels must be spaced far enough apart to avoid energy spillover between channels. A common scheme used to increase adjacent channel discrimination is to alternate between vertical and horizontal polarization. The isolation provided by crosspolarizing adjacent channels is on the order of ≥ 20 dB. At the edges of the band, a guard spacing is necessary to protect against interference into and from adjacent bands.

Another important consideration in radio-link design is RF interference, which may occur from sources internal or external to the radio system. The system designer should be aware of these interference sources in the area of each radio link, including their frequency, power, and directivity. Certain steps can be taken to minimize the effects of interference: good site selection, use of properly designed antennas and radios to reject interfering signals, and use of a properly designed frequency plan. A comprehensive frequency study is required in order to select frequencies that will not create or experience interference with existing systems operating in the same frequency bands. Access to a frequency database is required for interference analyses. The FCC prescribes the use of EIA/TIA Bulletin TSB10-F for interference analysis of microwave systems [22].

The effect of RF interference on a radio system depends on the level of the interfering signal and whether the interference is in an adjacent channel or is cochannel. A cochannel interferer has the same nominal radiofrequency as that of the desired channel. Cochannel interference arises from multiple use of the same frequency without proper isolation between links. Adjacent-channel interference results from the overlapping components of the transmitted spectrum in adjacent channels. Protection against this type of interference requires control of the transmitted spectrum, proper filtering within the receiver, and orthogonal polarization of adjacent channels.

The performance criteria for digital radio systems in the presence of interference are usually expressed in one of two ways: allowed degradation of the S/N (SNR) threshold or allowed BER. Both criteria are stated for a given signalto-interference ratio (S/I).

5. DIGITAL RADIO DESIGN

A block diagram of a digital radio transmitter and receiver is shown in Fig. 12. The traffic data streams at the input to the transmitter are usually in coded form, for example bipolar, and therefore require conversion to an NRZ (nonreturn-to-zero) signal with an associated timing signal. The multiplexer combines the traffic NRZ streams and any auxiliary channels used for orderwires into an aggregate data stream. This step is accomplished either by using pulse stuffing, which allows the radio clock rate to be independent of the traffic data, or by using a synchronous interface, which requires the radio and traffic data to be controlled by the same clock. The aggregate signal is scrambled to obtain a smooth radio spectrum and ensure recovery of the timing signal at the receiver. For phase modulation, some form of differential encoding is often employed to map the data into a change of phase from one signaling interval to the next. The modulator converts the digital baseband signals into a modulated intermediate frequency (IF), which is typically at 70 MHz when the final frequency is in the microwave band. The RF carrier is generated by a local oscillator, which is mixed with the IFmodulated signal to produce the microwave signal. The RF power amplification is accomplished by a traveling-wave tube (TWT) or a solid-state amplifier such as the gallium arsenide field-effect transistor (GaAs FET) amplifier. The final component of the transmitter is the RF filter, which shapes the transmitted spectrum and helps control the signal bandwidth.

At the receiver, the RF signal is filtered and then mixed with the local oscillator to produce an IF signal. The IF signal is filtered and amplified to provide a constant output level to the demodulator. Automatic gain control (AGC) in the IF amplifier provides variable gain to compensate for signal fading. Because the AGC voltage is a convenient indicator of received signal level, it is often used for performance monitoring or diversity combining. Fixed equalization is required to compensate for static amplitude or delay distortion from radio components, such as a TWT or filter, or to build out differential delay between RF channels in diversity operation. Adaptive equalization may also be required to deal with frequency-selective fading on the transmission path. Using the amplified and equalized IF signal, the demodulator recovers data and corresponding timing signals. Some type of performance monitoring is also commonly found in the demodulator, often based on eye pattern opening or pseudoerror techniques. The recovered baseband signal is next decoded and descrambled to reconstruct the aggregate datastream. The demultiplexer recovers the traffic datastreams and auxiliary channels. Finally, in the baseband encoder the standard data interface is generated.

5.1. Antennas

Microwave antennas used in line-of-sight transmission (terrestrial or satellite) provide high gain because radiated energy is focused within a small angular region. The power gain of a parabolic antenna is directly related to both the aperture area of the reflector and the frequency of operation. Relative to an isotropic antenna, the gain can be expressed as

$$G = \frac{4\pi \eta A}{\lambda^2} \tag{42}$$

where λ is the wavelength of the frequency, A is the actual area of the antenna in the same units as λ , and η is the efficiency of the antenna. The antenna efficiency is a number between 0 and 1 that reduces the theoretical gain because of physical limitations such as nonoptimum illumination by the feed system, reflector spillover, feed system blockage, and imperfections in the reflector surface. Converting (42) to more convenient units, the gain may be expressed in decibels as

$$G_{\rm dB} = 20\log f + 20\log d + 10\log \eta - 49.92 \tag{43}$$

where f is the frequency in megahertz and d is the antenna diameter in feet.

Waveguide or coaxial cable is required to connect the top of the radio rack to the antenna. Coaxial cable is limited to frequencies up to the 2-GHz band. Three types of waveguide are used-rectangular, elliptical, and circular-which vary in cost, ease of installation, attenuation, and polarization. The rectangular waveguide has a rigid construction, is available in standard lengths, and is typically used with short runs requiring only a single polarization. Standard bends, twists, and flexible sections are available, but multiple joints can cause reflections. The elliptical waveguide is semirigid, is available in any length, and is therefore easy to install. It accommodates only a single polarization, but has lower attenuation than does the rectangular waveguide. The circular waveguide has the lowest attenuation, about half that of the rectangular one, and also provides dual polarization and multiple band operation in a single waveguide. Its disadvantages are higher cost and more difficult installation.

5.2. Diversity Design

Diversity in LoS links is used to increase link availability by reducing the effects of multipath fading, improving the



a)

A typical arrangement for space diversity has two transmitters that operate on the same frequency and can be switched for output to a common antenna. One transmitter can operate in a hot standby mode, while the other is online; but as an alternative configuration, they can be combined to provide a 6-dB increase over the power available from a single transmitter. In this latter case, the failure of one transmitter power amplifier causes a 6-dB drop in output power but the link remains operational. The two receivers are connected to different antennas that are physically separated to provide the desired space diversity effect. The receiver outputs are fed to the combiner, which combines the two received signals. Space diversity, unlike frequency diversity, does not require an additional frequency assignment and is therefore more efficient in the use of spectrum. Its disadvantage is that additional antennas and waveguides are required, making it more expensive than frequency diversity arrangements.

In a typical arrangement for frequency diversity, two transmitters operate continuously on different frequencies but carry identical traffic. The receivers are connected to the same antenna but are tuned to separate frequencies. The combiner function is identical to that of the space diversity configuration. The use of frequency diversity doubles the spectrum amount required — a significant disadvantage in congested frequency bands. Unlike space diversity, however, frequency diversity provides two complete, independent paths, allowing testing of one path without interrupting service, while requiring only a single antenna per link end. Angle diversity is a newer form of diversity that has been shown to have cost and performance advantages over the more conventional diversity techniques already discussed. The basic idea behind angle diversity is that most deep fades, which are caused by two rays interfering with one another, can be mitigated by a small change in the amplitude or phase of the individual rays. Therefore, if a deep fade is observed on one beam of an angle diversity receiving antenna, switching to (or combining with) the second beam will change the amplitude or phase relationship and reduce the fade depth. Two techniques have been used to achieve angle diversity: a single antenna with two feeds that have a small angular displacement between them, or two separate antennas having a small angular difference in vertical elevation.

Variations of these protection arrangements include hot standby, hybrid diversity, and M:N protection. Hot standby arrangements apply to those cases where a second RF path is not deemed necessary, as with short paths where fading is not a significant problem. Here both pairs of transmitters and receivers are operated in a hot standby configuration to provide protection against equipment failure. The transmitter configuration is identical to that of space diversity. The receiving antenna feeds both receivers, tuned to the same frequency, through a power splitter. Hybrid diversity is provided by using frequency diversity but with the receivers connected to separate antennas that are spaced apart. This arrangement, which combines space and frequency diversity, improves the link availability beyond that realized with only one of these schemes.

For more efficient use of equipment and spectrum, diversity techniques are sometimes applied to a section of one or more links. In its simplest form, frequency diversity is used per section, with one protection channel used for N operational channels. This method can be extended to provide M: N protection, where M protection channels are shared by N operational channels. Further protection can be provided by using space diversity on a per-hop basis and frequency diversity on a section basis.

A diversity combiner performs the combining or selection of diversity signals. This function can be performed at RF [24], IF [25], or baseband [23]. Combiner techniques used in analog radio transmission [8] are generally applicable to digital radio. Phase alignment of the diversity signals becomes more important in digital radio systems, however, because of the potential occurrence of error bursts or loss of timing synchronization when combining misaligned diversity signals. Since delay equalization is simpler at baseband than at IF or RF baseband, "hitless" switching is a popular choice in digital radio combiners. This selection combiner uses some form of in-service performance monitor in each receiver to select the output signal after demodulation and data detection.

5.3. Adaptive Equalizer Design

Initial applications of wideband digital radios revealed that dispersion due to frequency-selective fading was the dominant source of multipath outages. These experiences led to the development and use of adaptive equalization so that outage requirements could be met. Since the introduction of the first adaptive equalizers to digital radio, these devices have undergone a rapid evolution. The degree of sophistication required of the equalizer depends primarily on the bit rate and pathlength. The types of equalizer in use today range from simple amplitude slope equalizers to complex transversal filter equalizers.

To obtain best equalizer performance, particularly for group delay distortion, transversal equalization techniques are now commonly applied to digital radio [26]. Figure 13 shows a block diagram of a QAM demodulator equipped with a transversal filter equalizer. The demodulator outputs are equalized by a forward baseband equalizer whose tap weights are controlled by decision feedback. The tap weights are continuously adjusted to correct channel distortion and provide the desired pulse response. The number of taps and the tap spacing in the transversal filters are design parameters that are chosen according to the system bit rate, channel fading characteristics, and radio outage requirements. Field tests of a baseband adaptive transversal filter applied to a 90-Mbps, 16-QAM radio show outage improvement by a factor of >3compared with the same radio equipped with an IF slope equalizer only [27].



Figure 13. QAM demodulator equipped with decision-directed transversal filter equalizer.

6. RADIO-LINK CALCULATIONS

Procedures for allocating radio-link performance must be based on end-to-end system performance requirements. Outages result from both equipment failure and propagation effects. Allocations for equipment and propagation outage are usually separated because of their different effects on the user and different remedy by the system designer. Here we will consider procedures for calculating values of key radio-link parameters, such as transmitter power, antenna size, and diversity design, based on propagation outage requirements alone. These procedures include the calculation of intermediate parameters such as system gain and fade margin. Automated design of digital LoS links through use of computer programs is now standard in the industry [6,17].

6.1. System Gain

System gain (G_s) is defined as the difference, in decibels, between the transmitter output power (P_t) and the minimum receiver signal level (RSL) required to meet a given bit error rate objective (RSL_m):

$$G_s = P_t - \text{RSL}_m \tag{44}$$

The minimum required RSL_m , also called receiver threshold, is determined by the receiver noise level and the signal-to-noise ratio required to meet the given BER. Noise power in a receiver is determined by the noise power spectral density (N_0) , the amplification of the noise introduced by the receiver itself (noise figure N_f), and the receiver bandwidth (B), The total noise power is then given by

$$P_N = N_0 B N_f \tag{45}$$

The source of noise power is thermal noise, which is determined solely by the temperature of the device. The thermal noise density is given by

$$N_0 = kT_0 \tag{46}$$

where k is Boltzmann's constant (1.38 \times 10^{-23} J/K) and T_0 is absolute temperature (K)

The reference for T_0 is normally assumed to be room temperature, 290 K, for which $kT_0 = -174$ dBm/Hz. The minimum required RSL may now be written as

$$RSL_m = P_N + SNR$$

$$= kT_0 BN_f + SNR$$
(47)

It is often more convenient to express (47) as a function of data rate *R* and E_b/N_0 . Since SNR = $(E_b/N_0)(R/B)$, we can rewrite (47) as

$$\mathrm{RSL}_m = kT_0 RN_f + \frac{E_b}{N_0} \tag{48}$$

The system gain may also be stated in terms of the gains and losses of the radio link:

$$G_s = L_p + F + L_t + L_m + L_b - G_t - G_r$$
(49)

where $G_s =$ system gain (dB)

- L_p = free-space path loss (dB), given by Eq. (4) F = fade margin (dB)
- $L_t = \text{transmission line loss from waveguide or}$ coaxials used to connect radio to antenna (dB)
- L_m = miscellaneous losses such as minor antenna misalignment, waveguide corrosion, and increase in receiver noise figure due to aging (dB)
- $L_b =$ branching loss due to filter and circulator used to combine or split transmitter and receiver signals in a single antenna
- $G_t = \text{gain of transmitting antenna}$

 G_r = gain of receiving antenna

The system gain is a useful figure of merit in comparing digital radio equipment. High system gain is desirable since it facilitates link design—for example, by easing

6.2. Fade Margin

F

The traditional definition of fade margin is the difference, in decibels, between the nominal RSL and the threshold RSL as illustrated in Fig. 6. An expression for the fade margin F required to meet allowed outage probability P(o)may be derived from (29) for an unprotected link, from (37) for a space diversity link, and from (38) for a frequency diversity link, with the following results:

$$F = 30\log D + 10\log(abf) \tag{50}$$

 $-56 - 10 \log P(o)$ (unprotected link)

$$F = 20\log D - 10\log S \tag{51}$$

 $+5\log(ab) - 7.2 - 5\log P(o)$ (space diversity link)

$$= 20 \log D + 15 \log f + 5 \log(ab) - 5 \log(\Delta f)$$
(52)

 $-36.5 - 5 \log P(o)$ (frequency diversity link)

This definition of fade margin has traditionally been used to describe the effects of fading at a single frequency for radio systems that are unaffected by frequency-selective fading or to radio links during periods of flat fading. For wideband digital radio without adaptive equalization; however, dispersive fading is a significant contributor to outages so that the flat fade margins do not provide a good estimate of outage and are therefore insufficient for digital link design. Several authors have introduced the concept of effective, net, or composite fade margin [18] to account for dispersive fading. The effective (or net or *composite*) *fade margin* is defined as that fade depth that has the same probability as the observed probability of outage. The difference between the effective fade margin measured on the radio link and the flat fade margin measured with an attenuator is then an indication of the effects of dispersive fading. Since digital radio outage is usually referenced to a threshold error rate, BER_t , the effective fade margin (EFM) can be obtained from the relationship

$$P(A \ge \text{EFM}) = P(\text{BER} \ge \text{BER}_t) \tag{53}$$

where A is the fade depth of the carrier. The results of Eqs. (50)-(52) can now be interpreted as yielding the effective fade margin for a probability of outage given by the right-hand side of (53). Conversely, note that Eqs. (29), (37), and (38) are now to be interpreted with F equal to the EFM.

The effective fade margin is derived from the addition of up to three individual fade margins that correspond to the effects of flat fading, dispersion, and interference:

$$EFM = -10 \log(10^{-FFM/10} + 10^{-DFM/10} + 10^{-IFM/10})$$
 (54)

Here the flat fade margin (FFM) is given as the difference in decibels between the unfaded signal-to-noise ratio $(SNR)_u$ and the minimum signal to-noise ratio $(SNR)_m$ to meet the error rate objective, or

$$FFM = (SNR)_u - (SNR)_m \tag{55}$$

Similarly, we define the dispersive fade margin (DFM) and interference fade margin (IFM) as

$$DFM = (SNR)_d - (SNR)_m$$
(56)

and

$$IFM = (S/I) - (SNR)_m$$
(57)

where $(SNR)_d$ represents the effective noise due to dispersion and (S/I) represents the critical S/I below which the BER is greater than the threshold BER. Each of the individual fade margins can also be calculated as a function of the other fade margins by using (54), as in

$$FFM = -10\log(10^{-EFM/10} + 10^{-DFM/10} + 10^{-IFM/10})$$
(58)

BIOGRAPHY

David R. Smith is a professor in the Department of Electrical and Computer Engineering at The George Washington University, Washington, D.C. He received a D.Sc. in electrical engineering and computer science from The George Washington University in 1977; an M.S.E.E. from Georgia Tech, Atlanta, in 1970; and a B.S in physics from Randolph-Macon College, Randolph, Virginia, in 1967. He has worked as an electrical engineer and manager within the Department of Defense, Washington, D.C., for both the Navy Department and the Defense Information Systems Agency. Since 1967 he has held adjunct, visiting, research, and regular faculty positions at Georgia Tech, The George Washington University, and George Mason University, Fairfax, Virginia. He has also consulted with numerous companies in the areas of wireless and digital communications. His publications include over 20 IEEE articles, the books Digital Transmission Systems published by Kluwer and Emerging Public Safety Wireless Communication Systems published by Artech House, and chapters contributed to two other books. Current interests include research in areas of wireless telecommunications, advanced networking, digital communications, propagation modeling, and computer simulation and modeling of communication systems.

BIBLIOGRAPHY

- 1. Henry R. Reed and Carl M. Rusell, *Ultra High Frequency Propagation*, Chapman & Hall, London, 1966.
- K. Bullington, Radio propagation fundamentals, Bell Syst. Tech. J. 36: 593-626 (May 1957).
- J. H. Van Vleck, Radiation Laboratory Report 664, MIT, Cambridge, MA, 1945.
- 4. Reports of the ITU-R, 1990, Annex to Vol. V, *Propagation in Non-Ionized Media*, ITU, Geneva, 1990.
- S. H. Lin, Nationwide long-term rain rate statistics and empirical calculation of II-GHt microwave rain attenuation, *Bell Syst. Tech. J.* 56: 1581-1604 (Nov. 1977).

2572 TEST AND MEASUREMENT OF OPTICALLY BASED HIGH-SPEED DIGITAL COMMUNICATIONS SYSTEMS AND COMPONENTS

- D. R. Smith, A computer-aided design methodology for digital line-of-sight links, Proc. 1990 Int. Conf. Communications, 1990, pp. 59–65.
- 7. Engineering Considerations for Microwave Communications Systems GTE Lenkurt Inc., San Carlos, CA, 1981.
- 8. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill, New York, 1966.
- W. T. Barnett, Multipath propagation at 4, 6, and 11 GHz, Bell Syst. Tech. J. S1: 321–361 (Feb. 1972).
- A. Vigants, Space diversity performance as a function of antenna separation, *IEEE Trans. Commun. Technol.* COM-16(6): 831-836 (Dec. 1968).
- A. Vigants and M. V. Pursley, Transmission unavailability of frequency diversity protected microwave FM radio systems caused by multipath fading, *Bell Syst. Tech. J.* 58: 1779–1796 (Oct. 1979).
- L. C. Greenstein and B. A. Czekaj, A polynomial model for multipath fading channel responses, *Bell Syst. Tech. J.* 59: 1197-1205 (Sept. 1980).
- W. D. Rummler, A new selective fading model: Application to propagation data, *Bell Syst. Tech. J.* 58: 1037–1071 (May/June 1979).
- W. C. Jakes, Jr., An approximate method to estimate an upper bound on the effect of multipath delay distortion on digital transmission, *Proc. 1978 Int. Conf. Communications*, 1978, pp. 47.1.1-47.1.5.
- M. Emshwiller, Characterization of the performance of PSK digital radio transmission in the presence of multipath fading, *Proc. 1978 Int. Conf. Communications*, 1978, pp. 47.3.1-47.3.6.
- C. W Lundgren and W. D. Rummler, Digital radio outage due to selective fading—observation vs. prediction from laboratory simulation, *Bell Syst. Tech. J.* 58: 1073-1100 (May/June 1979).
- T. C. Lee and S. H. Lin, The DRDIV computer program—a new tool for engineering digital radio routes, 1986 GLOBE-COM, pp. 51.5.1–51.5.8.
- C. W. Anderson, S. G. Barber, and R. N. Patel, The effect of selective fading on digital radio, *IEEE Trans. Commun.* COM-27(12): 1870-1876 (Dec. 1979).
- T. S. Giuffrida, Measurements of the effects of propagation on digital radio systems equipped with space diversity and adaptive equalization, *Proc. 1979 Int. Conf. Communications*, 1979, pp. 48.1.1-48.1.6.
- D. R. Smith and J. J. Cormack, Improvement in digital radio due to space diversity and adaptive equalization, *Proc. 1984 Global Telecommunications Conf.*, 1984, pp. 45.6.1-45.6.6.
- Federal Communications Commission Rules and Regulations, Part 101, Fixed Microwave Services, Aug. 1, 1996.
- 22. TIA/EIA Telecommunications Systems Bulletin, TSB10-F, Interference Criteria for Microwave Systems, June 1994.
- C. M. Thomas, J. E. Alexander, and E. W. Rahneberg, A new generation of digital microwave radios for U.S. military telephone networks, *IEEE Trans. Commun.* COM-27(12): 1916-1928 (Dec. 1979).
- 24. 1. Horikawa, Y. Okamoto, and K. Morita, Characteristics of a high capacity 16 QAM digital radio system on a multipath fading channel, *Proc. 1979 Int. Conf. Communications*, 1979, pp. 48.4.1-48.4.6.

- G. deWitte, DRS-8: System design of a long haul 91 Mb/s digital radio, Proc. 1978 Natl. Telecommunications Conf., 1978, pp. 38.1.1-38.1.6.
- C. A. Siller, Jr., Multipath propagation, *IEEE Commun. Mag.* 6–15 (Feb. 1984).
- 27. G. L. Fenderson, S. R. Shepard, and M. A. Skinner, Adaptive transversal equalizer for 90 Mbls 16-QAM systems in the presence of multipath propagation, *Proc. 1983 Int. Conf. Communications*, 1983, pp. C8.7.1–C8.7.6.

TEST AND MEASUREMENT OF OPTICALLY BASED HIGH-SPEED DIGITAL COMMUNICATIONS SYSTEMS AND COMPONENTS

Greg D. LeCheminant Agilent Technologies Santa Rosa, California

Gauging the performance of a communications system can be accomplished in a number of ways and with a variety of parameters. The fundamental task of any communications system is to reliably transport information. Thus the one parameter that can describe the overall health of a digital communications system is the bit error-rate (BER). A BER test is a direct measure of the number of bits received incorrectly compared to the total number of bits transmitted. Prior to integration into a working system, it is also important to characterize the performance of the individual components and building blocks that make up the system. Rather than "system level" performance parameters such as BER, tests that describe physical or parametric properties can be used. Examples include signal strength, noise, spectral content, and tolerance to jitter.

1. MEASURING BIT ERROR RATIO

Modern optically based communications systems perform extremely well. BERs on the order of 1 bit received in error per trillion bits transmitted are common. This is often expressed as a BER of 1×10^{-12} . It is not uncommon for this parameter to be incorrectly referred to as the bit error *rate* instead of bit error ratio. However, the 1×10^{-12} number mentioned above does not indicate the rate at which bits are received in error. This would require knowing both the transmission rate in bits per second and the BER (ratio of errored bits to transmitted bits).

Measuring BER is typically achieved using a BER tester or "BERT." A BERT consists of two major building blocks. First there is a pattern generator. The pattern generator performs two major functions. It determines the specific data sequence that the system under test will transmit. This is commonly referred to as the *test pattern*. It is essential that specific, known patterns be used so that it can be determined whether the system under test correctly transported the data. The pattern generators also can provide control over characteristics of the test signal such as waveform amplitude and to some degree the pulseshape. The second block of the BERT is the error



Figure 1. BERT test set.

detector. The error detector is used to verify that each bit sent through the system under test has been correctly received.

The process of performing a BER measurement is as follows (see Fig. 1):

- 1. The pattern generator is configured to produce a specific data sequence, which is fed to the system under test.
- 2. The system transmits the pattern to its receiver. The receiver must determine the logic level for each incoming bit. In most cases, the signal at the receiver decision circuit will have been degraded because of attenuation and dispersion in the transmission path. The system receiver effectively produces a regenerated output signal or pattern according to what it thought the incoming bits were. A perfect receiver would never make a mistake in assessing the logic level of the input signal.
- 3. A separate pattern generator within the error detector section of the BERT produces a pattern identical to that originally fed to the system under test.
- 4. The error detector pattern and the signal or pattern from the system receiver are aligned in time and compared bit for bit. Whenever there is disagreement, an error is counted.
- 5. The BER is determined from the number of bits found in error and the total number of bits checked.

An error detector can be viewed as a decision circuit followed by an "exclusive OR" (XOR) logic circuit. Data from the system or circuit under test are fed to the error detector. A clock signal is also fed to the error detector. This clock signal is often a clock signal that the receiver of the system or circuit under test has derived from the data being transmitted. The clock is used to time the decision circuit and XOR gate. The timing of the decision circuit must be optimized within one clock cycle to sample the data signal at the ideal point in time (usually the middle of the bit period). Thus a "clock to data" alignment must be performed. This is usually an automatic adjustment available in the BERT. There is also an optimum amplitude threshold the error detector decision circuit will use to determine whether incoming data are at the 1 or 0 level. A BERT will also have an alignment procedure to determine this optimum level. The third basic step in setting up a BER test is to align the internal pattern generator of the error detector with the pattern from the system or circuit under test so that the xor function is being performed on identical locations in each pattern. This too is an automatic procedure available in all modern BERTs.

Although a pattern generator can produce virtually any pattern sequence desired, most BER measurements are performed using pseudorandom binary sequences (PRBSs). A PRBS is a data pattern that attempts to replicate truly random data yet is completely deterministic (a requirement for performing a BER measurement). PRBS patterns have lengths of $2^N - 1$. For example, a $2^7 - 1$ pattern is 127 bits in length and will include all sequences 7 bits in length with the exception of 7 sequential 0's. A $2^{10} - 1$ PRBS is 1023 bits in length and has all sequences 10 bits in length with the exception of 10 sequential 0's (zeros). Pattern generators commonly produce pattern lengths of up to $2^{31} - 1$.

PRBS patterns are generated using a series of shift registers with feedback taps (see Fig. 2). One of the unique properties of a PRBS is that when compared to itself, the BER will be 50% except when the patterns from the system under test and the error detector are exactly aligned. This allows the alignment process to be performed quickly and efficiently.

There are several beneficial properties of PRBS patterns. The spectral content of a PRBS is quite broad; thus it can expose resonances in test devices. PRBS patterns can be decimated to produce shorter PRBS patterns. This is useful when testing multiplexing and demultiplexing circuits. (For example, the shorter patterns produced by a 1-to-4 demultiplexer are still PRBSs and can thus be easily tested for BER.) PRBS patterns can also



Sequence length	Shift-register configuration		
2 ⁷ –1	$D^7 + D^6 + 1 = 0$, inverted		
2 ¹⁰ –1	$D^{10} + D^7 + 1 = 0$, inverted		
2 ¹⁵ -1	$D^{15} + D^{14} + 1 = 0$, inverted		
2 ²³ –1	$D^{23} + D^{18} + 1 = 0$, inverted		
2 ³¹ -1	$D^{31} + D^{29} + 1 = 0$, inverted		

Figure 2. Generating the PRBS.

be generated at extremely high speeds, limited only by the hardware generating the pattern.

In addition to PRBS patterns, specialized patterns can also be created and generated with a BERT. A special pattern might be created that causes a transmitter to produce more jitter than if a PRBS were used. Communications protocols often require specific bit sequences to "frame" the data payload. Thus a pattern may be built to include the framing bits. These patterns are not produced by a specific circuit topology, but instead are loaded into the memory of the pattern generator. The pattern generator then produces a data sequence according to the loaded pattern. (The internal pattern generator of the error detector can perform similarly.)

While the BER result is indicative of the system performance, it does not provide any indication of underlying reasons for poor or marginal performance. Sometimes the way in which errors occur can provide insight into the causes for errors. Consider the case when 2 or 3 adjacent bits are received in error. If the overall BER is 1×10^{-9} , the probability of 2 bits being received in error as a result of random mechanisms is one in 1×10^{18} . For three to occur in sequence the probability would be one in 1×10^{27} . Because 10^{27} bits at 1 Gbps (gigabits per second) takes millions of years to transmit, 3 sequential bits being in error due to random causes is extremely unlikely in this case. It is far more likely that these errors would be due to something deterministic.

Several techniques can help troubleshoot the root causes for BER. Examples include examining the intervals (time or bits) between errors, or whether errors occur mainly for logic 1s or logic 0s. If errors occur separated by some multiple of 10 bits, and there is a 10-bit-wide parallel bus somewhere in the system, there could be something physically wrong with one of the bus lines.

With the common occurrence of BERs of 1×10^{-12} and even lower in today's high-speed telecommunications systems, what should be the expected time required to verify that a system is operating at a low BER? It is ideal to collect 100 or even 1000 errors to be confident that the error ratio performance is being achieved. However, at an error ratio of 1×10^{-12} and a data rate on the order of 1 Gbps, it will take $11\frac{1}{2}$ days to collect 1000 errors! If the data rate is increased to 10 Gbps and the number of errors collected is reduced to only 100, it will still take several hours to complete the measurement.

Certain techniques can be used to try to determine BER in a reduced timeframe. Both techniques to be discussed rely on making BER measurements in nonideal conditions, thus increasing the BER and extrapolation to determine the ideal BER. One technique is to stress the signal being tested. The other technique is to adjust the sampling threshold of the error detector to nonideal levels. A common technique to stress a signal is to simply attenuate it prior to reaching the system receiver. As the signal level is reduced, the likelihood that a receiver will incorrectly interpret 1s as 0s and 0s as 1s increases. The BER will be artificially increased and then take less time to measure. By measuring the BER at several levels of attenuation, a BER-received power curve can be plotted (see Fig. 3). Extrapolation of the curve can then be used to determine the BER with no external signal attenuation. Extrapolation requires an assumption that the dominant error mechanisms are the same at both high and low signal levels. This usually implies random noise. However, it is possible that a deterministic but extremely low-probability mechanism exists and is not seen except at high power levels where noise is less likely to cause errors. Thus final system qualification may require a lengthy BER measurement at full signal power.

The BER will also be artificially increased if the error detector sampling threshold is increased or decreased from the ideal. However, this type of BER analysis is typically performed on the signal at the test receiver input rather than its output. The BER analysis is not made on the complete system, but instead is an assessment of the quality of the signal being presented to the receiver. This is called a Q factor measurement. The Q factor is essentially the signal-to-noise ratio (SNR) of the signal. If the dominant mechanism for errors in a system is random noise, then BER can be estimated by the Q factor parameter.



Figure 3. BER versus received power.

The measurement process is as follows. The transmitted signal from the system under test is injected directly into the error detector. (If the system is optical, a linear optical to electrical converter is used to create an electrical input to the error detector.) BER measurements are made with the error detector sampling threshold at several levels above and below the ideal threshold. From each BER result, a signal-to-noise parameter or Q value can be estimated. Again, these measurements are made with sampling thresholds that yield BERs worse than 1×10^{-9} and thus can be performed quickly. With several BER values collected, the BERs are converted to Q values and plotted against sampling threshold. From this plot, the optimum Q factor and optimum sampling threshold can be extrapolated (see Fig. 4).

What does the optimum Q factor tell us? Just as Q factor can be derived from BER, the optimum BER can be derived from the optimum Q factor. The optimum Q factor indicates the highest level of system performance that can be achieved with an ideal decision circuit as a receiver. This is useful for characterizing ultra-low-BER systems that would require extremely long test times for actual BER verification. It is important to recognize the assumptions that are made. The most important one is that the dominant error mechanism is random noise. The mathematical Q/BER relationships are based on this. In the end, final qualification of a system will likely require basic BER verification.

2. WAVEFORM ANALYSIS

Analysis of waveforms can yield a wealth of information about the overall quality of a high-speed digital transmitter or the transmission section of a communication system. In the R&D lab, a simple visual inspection of the signal allows a designer to quickly assess basic performance. In a manufacturing environment, several key parameters can



Figure 4. Estimating low BERs through *Q*-factor measurements.

be derived from a transmitter's waveform. Waveforms are viewed on oscilloscopes that display signal amplitude as a function of time. There are two basic formats to view a digital communications signal, either as a pulsetrain or as an eye diagram.

A *pulsetrain* is a display of some segment or sequence of the communications signal. For example, if a 1-Gbps signal were being examined, each bit is 1 ns (nanosecond) in duration. If the oscilloscope timebase were set to a width of 10 ns, 10 sequential bits could be displayed. Determining which 10 bits from a long data sequence are displayed is a function of how the oscilloscope is triggered. Typically, a pattern generator will produce a "pattern trigger." This is a pulse produced at the beginning of each repetition of the data pattern, such as a PRBS. Triggering the oscilloscope with the pattern trigger and adjusting the oscilloscope time delay can display any specific section of the pattern (see Fig. 5). Very long patterns present some difficulty in displaying pulsetrains. Because the trigger pulse is generated only once per every repetition of the pattern, the time between trigger events can become very large. For wide bandwidth sampling oscilloscopes, one data point is sampled for every trigger event. If a waveform is composed of 1000 sample points, the entire pattern must be transmitted 1000 times to complete acquisition of the waveform. For long pattern lengths it can take several minutes and even hours to produce a single waveform. Thus pulsetrains are usually displayed only when examining relatively short patterns.

Another difficulty when examining pulsetrains is that only a few bits can be examined at a given time. More bits can be displayed by decreasing the resolution of the oscilloscope timebase, but important details are usually lost because of this reduced resolution. Often when examining a high-speed digital communication signal it is desirable to determine the overall performance of the system for all patterns of data. It would be ideal to see this in one simple display. This can be achieved through the eye diagram. The *eye diagram* is a composite display of waveform samples acquired throughout the entire data pattern displayed on a common timebase. Consider the eight waveforms that can be generated from a 3-bit sequence $(000, 001, \ldots, 110, 111)$. If these eight waveforms are all placed on a common amplitude-time grid, the eye diagram is displayed (see Fig. 6).

With an eye diagram it is difficult to view the waveform from any individual bit. However, much information is available regarding the overall performance of the



Figure 6. Building the eye diagram.



Figure 5. Pulsetrain waveform of a digital communications signal.

transmitted waveform. If the eye diagram begins to close horizontally, this is an indication of excessive waveform timing jitter. Slow rise and fall times cause vertical eye closure. Eye closure due to any mechanism presents a significant system level problem because it makes the decision process more difficult for the receiver at the end of the communication system.

An eye diagram is created when the oscilloscope is triggered with a clock signal that is synchronous to the data. In contrast to triggering with a pattern trigger, triggering with a clock signal allows the oscilloscope to acquire samples throughout the data pattern. Divided clocks, such as a rate $\frac{1}{4}$ or rate $\frac{1}{16}$ are also acceptable. The requirement is that the divided clock be synchronous with the data, and that the divisor be an integer. The oscilloscope can trigger on the data signal itself and create an eye diagram. However, although the displayed eye diagram may appear to be complete, approximately 75% of the data will be missing form the eye diagram. This is because only one of the four combinations of 2 adjacent bits (e.g., the 0–1 combination) will produce a signal transition that the oscilloscope can trigger on. Thus triggering on the data itself should be avoided whenever possible.

In an R&D environment many engineers and scientists have learned how to quickly gauge the quality of a signal through quick visual inspection of the eye diagram. In the most basic sense, an "open" eye diagram is indicative of a quality signal, while a "closed" eye is indicative of signal impairments.

In a manufacturing environment, the eye diagram is used to obtain specific parameters that indicate performance of high-speed digital transmitters. One common example is the measurement of the extinction ratio. The *extinction ratio* is used to describer how efficiently an optical transmitter converts its available signal strength to modulation power. In mathematical terms, it is simply the ratio of the power in a logic level one (1) to the power in a logic level zero (0). Since the eye diagram is composed of a multitude of logic ones and logic zeros, a statistical analysis is performed to determine the aggregate logic one power and the aggregate logic zero. This is achieved through the use of histograms. First a slice of data is acquired for the upper central portion of the eye diagram. A vertical histogram is constructed from these data. The mean of this histogram represents the power level of the aggregate ones making up the eye diagram. A similar process is used in the lower central portion of the eye diagram to determine the power of the aggregate logic zero (see Fig. 7). (A high extinction ratio is typically achieved by using very little power to transmit zeros. When this is achieved, virtually all of the available laser power is being used to transmit information. Thus, a high extinction ratio is indicative of a highly efficient use of laser power.)

The extinction ratio is just one example of histogram based statistical analysis to derive specific parameters of the eye diagram. Other measurements include optical modulation amplitude (OMA), rise time, fall time, jitter, and signal-to-noise-ratio. Again, these measurements are not made on individual bits, but rather on the composite eye diagram, thus yielding the overall performance of the transmitter with a single measurement.

Recall that it is desirable to have an open eye diagram in both the vertical and horizontal sense. It is difficult to measure a simple numerical parameter that can describe the openness of the eye diagram. Instead, a process called "eye-mask testing" is used. An eye-mask is a constellation of solid polygons that represent where the eye diagram waveform may *not* exist. A typical eye-mask consists of a polygon located in the center of the eye diagram, as well as



Figure 7. Measuring extinction ratio.



Figure 8. Eye-mask testing.

one polygon above and one polygon below (see Fig. 8). The eye diagram is not allowed to intersect or "violate" any of the mask polygons. The minimum acceptable opening for the eye diagram is then defined by the size and shape of the central polygon.

The shape of the displayed eye diagram can be altered by the frequency response of the oscilloscope measurement channel. It is common for directly modulated high-speed communication lasers to exhibit significant overshoot and ringing during the transition from a low-power logic 0 to a higher-power logic 1. It takes a widebandwidth oscilloscope channel to view this phenomenon. For example, if a laser is transmitting 2.5-Gbps data, the oscilloscope bandwidth needs to approach 10 GHz or higher for an accurate representation of the true waveform.

Eye-mask testing is a key element of most industry standards that specify high-speed optically based communication systems. To achieve consistent results across the industry, it is essential to specify the frequency response of the measurement channel. Thus in addition to defining the shape of the eye-mask, the measurement system is also defined through the concept of a reference receiver. A reference receiver usually consists of a photodetector followed by a lowpass filter (see Fig. 9). The combined response of the two elements typically follows a fourthorder Bessel-Thomson frequency response. This response is chosen since it closely approximates the response of a Gaussian filter. A Gaussian response yields minimal distortion of the waveform.

It is interesting to note the bandwidth that is normally specified for a reference receiver. In most communication standards, the -3-dB bandwidth is set to be 75% of the optical bit rate. For example, a reference receiver for a 10-Gbps system would have a bandwidth of 7.5 GHz. Initially



Figure 9. Eye-mask reference receiver.

this seems counterintuitive. Intentionally reducing the measurement bandwidth is likely to change the shape of the eye diagram. Effects such as the overshoot and ringing mentioned above can literally disappear when the bandwidth is reduced. The reduced bandwidth of the reference receiver can actually yield a waveform that will pass the eye-mask test that would otherwise fail with a wider-bandwidth receiver.

To understand the logic behind this, consider the main intent of the test. The usability of the transmitter in a real communications system needs to be verified. This objective is very different from trying to produce the most accurate image of the waveform shape. Most communication systems have receivers with bandwidths just wide enough to allow accurate determination of input signal levels. If the receiver bandwidth is wide, internal noise will increase, and likely degrade system BER. If the receiver bandwidth is too low, a signal making a 0-to-1 transition will be sluggish and not reach full amplitude within the bit period. Somewhere in between these extremes is an ideal receiver bandwidth. Although communication system receivers are not normally specified to have a "75% of bit rate" bandwidth, they do not normally have relatively wide bandwidths. Thus the reference receiver measurement approach has proved to be an effective way to verify transmitter performance.

BIOGRAPHY

Greg D. LeCheminant received the B.S. degree in 1983 in Electronics Engineering Technology and M.S. degree in 1984 in Electrical Engineering from Brigham Young University in Provo, Utah. He joined the Hewlett-Packard company in 1985 as a manufacturing development engineer working in the production of microwave subassemblies for high-performance signal generators. In 1989 he accepted a marketing position for development of instrumentation used for high-speed optoelectric device and component characterization. He is currently employed by Agilent Technologies (Formerly Hewlett-Packard Test and Measurement) involved in the development of communications industry standards and measurement techniques and applications in high-speed digital communications.

THRESHOLD DECODING

WILLIAM W. WU JAMES L. MASSEY Consultare Technology Group Bethesda, Denmark

1. INTRODUCTION

Threshold decoding was one of the earliest practical techniques introduced for the decoding of linear errorcorrecting codes (or "parity-check codes"). Before explaining threshold decoding in general, we give two simple examples to illustrate its main features. Throughout this article, we consider only binary codes — partly for simplicity but also because threshold decoding is not well suited to the decoding of nonbinary codes.

Example 1. Consider the encoder for the binary linear (n = 6, k = 3) code in which the length n = 6 binary code word $\mathbf{v} = [v_1 \ v_2 \ v_3 \ v_4 \ v_5 \ v_6]$ is determined from the length k = 3 binary information sequence $\mathbf{u} = [u_1 \ u_2 \ u_3]$ by

 $[v_1 \, v_2 \, v_3 \, v_4 \, v_5 \, v_6] = [u_1 \, u_2 \, u_3] \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$

where binary arithmetic, that is, arithmetic modulo two in which $1 \oplus 1 = 0$, is understood. The above 3×6 matrix is called a *generator matrix* for the code. This generator matrix **G** specifies a *systematic encoder* in the sense that the information bits appear unchanged within the code word, viz. in the first k = 3 positions, as follows

from the fact that $\mathbf{G} = [\mathbf{I}_3 \\ \vdots \\ \mathbf{P}]$ where \mathbf{I}_k denotes the $k \times k$ identity matrix. From this matrix equation, we see that $v_1 = u_1$, that $v_3 \oplus v_5 = u_3 \oplus (u_1 \oplus u_3) = u_1$, and that $v_2 \oplus v_6 = u_2 \oplus (u_1 \oplus u_2) = u_1$. One says that these three sums of encoded bits are *orthogonal on the information*

bit u_1 in the sense that each sum is equal to u_1 plus one or more encoded bits, but no encoded bit appears in more than one of the sums. Suppose now that the code word **v** is transmitted over a binary symmetric channel (BSC) with crossover probability *p* where 0 . Then thelength n = 6 binary received word **r** can be written as $\mathbf{r} = [r_1 r_2 r_3 r_4 r_5 r_6] = \mathbf{v} \oplus \mathbf{e}$ where $\mathbf{e} = [e_1 e_2 e_3 e_4 e_5 e_6]$ is the binary error pattern, each of whose bits independently has probability *p* of being 1. Because $r_i = v_i \oplus e_i \neq v_i$ if and only if $e_i = 1$, the Hamming weight of **e**, that is, the number of its nonzero components, is the actual number of errors that occurred during the transmission of \mathbf{v} over the BSC. If we now form the above three sums orthogonal on u_1 using the received bits in place of the transmitted bits, we obtain $r_1 = v_1 \oplus e_1 = u_1 \oplus e_1$, $r_3 \oplus r_5 = v_3 \oplus e_3 \oplus v_5 \oplus e_5 =$ $u_1 \oplus e_3 \oplus e_5$, and $r_2 \oplus r_6 = v_2 \oplus e_2 \oplus v_6 \oplus e_6 = u_1 \oplus e_2 \oplus e_6$. These three sums of received bits are orthogonal on the information bit u_1 in the sense that each sum is equal to u_1 plus one or more error bits, but no error bit appears in (or "corrupts") more than one of these sums. It follows that if there is at most one actual error, that is, if at most one of the error bits is a 1, then the information bit u_1 can be correctly found at the receiver as the majority vote of r_1 , $r_3 \oplus r_5$, and $r_2 \oplus r_6$. Entirely similar arguments show that, again if at most one of the error bits is a 1, the information bit u_2 can also be correctly found by taking the majority vote of r_2 , $r_3 \oplus r_4$, and $r_1 \oplus r_6$ and that the information bit u_3 can be correctly found by the majority vote of r_3 , $r_1 \oplus r_5$, and $r_2 \oplus r_4$. This manner of decoding is an example of majority decoding, the earliest and simplest form of "threshold decoding." Because majority decoding in this example corrects all single errors, the minimum distance d_{\min} of the code must be at least 3. That the minimum distance is exactly three can be seen from the fact that the information sequence $\mathbf{u} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ gives the code word $\mathbf{v} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \end{bmatrix}$ with Hamming weight 3, that is, at distance 3 from the all-zero code word.

From this example, one deduces that if, for each information bit in a binary linear code, a set of δ sums of encoded bits orthogonal on this bit can be formed, then all patterns of $(\delta - 1)/2$ or fewer errors can be corrected by majority decoding. This implies that the minimum distance d_{\min} of the code is at least δ . One says that the code can be *completely orthogonalized* if $d_{\min} = \delta$.

Example 2. Consider the encoder for the (7, 4) binary cyclic code in which the code word $\mathbf{v} = [v_1 v_2 v_3 v_4 v_5 v_6 v_7]$ is determined from the information sequence $\mathbf{u} = [u_1 u_2 u_3 u_4]$ by

$$\begin{bmatrix} v_1 v_2 v_3 v_4 v_5 v_6 v_7 \end{bmatrix} = \begin{bmatrix} u_1 u_2 u_3 u_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

This is the (7, 4) Hamming code with $d_{\min} = 3$. It is easy to verify that it is impossible to form three sums of encoded bits that are orthogonal on any one of the four information bits. However, we note that $v_2 \oplus v_3 = u_2 \oplus u_3$, $v_1 \oplus v_6 = u_2 \oplus u_3$ and $v_4 \oplus v_7 = u_2 \oplus u_3$ so that one can form three sums of encoded bits that are orthogonal on the sum $u_2 \oplus u_3$ of information bits. If the code word \mathbf{v} is transmitted over a BSC, majority decoding of the three corresponding sums of received bits will determine the sum $u_2 \oplus u_3$ correctly when at most one actual error occurs in transmission. Let $(u_2 \oplus u_3)^{\Delta}$ denote the decoding decision for $u_2 \oplus u_3$. Then $v_6 \oplus (u_2 \oplus u_3)^{\Delta} = u_1 \oplus (u_2 \oplus u_3) \oplus (u_2 \oplus u_3)^{\Delta}$ is equal to u_1 when the decoding decision is correct. It follows that $v_1 =$ $u_1, v_3 \oplus v_4 \oplus v_5 = u_1$ and $v_6 \oplus (u_2 \oplus u_3)^{\Delta} = u_1$ constitute a set of three sums of encoded bits and previously decoded *bits* that are orthogonal on the information bit u_1 when the previous decoding decision is correct. Hence, u_1 can now be determined correctly by majority decoding when at most one actual error occurs in transmission. Because the code is cyclic, the information bit u_2 can similarly be determined simply by increasing the indices cyclically (i.e., increasing n = 7 gives 1) in the expression used to determine u_1 , and similarly for u_3 and u_4 . Thus, majority decoding of this code corrects all single errors. This is in fact complete minimum-distance decoding because, in a Hamming code, every received word is at distance at most 1 from some code word.

This decoding of a Hamming code exemplifies majority decoding with L-step orthogonalization for L = 2. The number of steps is the number of levels of decoding decisions (all of which use at least δ orthogonal sums) required to determine an information bit. We can deduce from this example that if, for each information bit in a binary linear code, a set of δ sums of encoded bits and previously decoded bits orthogonalization, then all patterns of $(\delta - 1)/2$ or fewer errors can be corrected by majority decoding and hence the minimum distance d_{\min} of the code is at least δ . One says that the code can be L-step orthogonalized if $d_{\min} = \delta$. Note that 1-step orthogonalization is the same as complete orthogonalization as defined above.

2. EARLY HISTORY

The decoding algorithm given by Reed [1] in 1954 for the binary linear codes found two years earlier by Muller, which are now universally called the Reed-Muller codes, was the first application of majority decoding to multiple-error-correcting codes. The μ^{th} order Reed-Muller code of length $n = 2^m$, where $0 \le \mu < m$, has $k = \sum_{i=0}^{\mu} \binom{m}{i}$ information bits and minimum distance $d_{\min} = 2^{m-1}$. Reed showed that this code can be $(\mu + 1)$ -step orthogonalized (although this terminology did not come into use until eight years later).

Yale [2] and Zierler [3] independently showed in 1958 that the binary maximal-length codes could be completely orthogonalized. Mitchell et al. [4] made an extensive study of majority decoding of binary cyclic codes in 1961 that included majority decoding algorithms for the Hamming codes, for the $(n = 73, k = 45, d_{\min} = 9)$ and the $(n = 21, k = 11, d_{\min} = 5)$ cyclic codes found by Prange [5], and for the $(n = 15, k = 7, d_{\min} = 5)$ Bose-Chaudhuri-Hocquenghem (BCH) code.

In his 1962 doctoral thesis at M.I.T., which appeared essentially unchanged as a monograph [6] the following year, Massey formulated threshold decoding as a general technique comprising both majority decoding and also what he called *APP decoding*, where "APP" is short for "a posteriori probability."

In APP decoding, one again uses orthogonal sums on some bit (or sum of bits), but now one takes into account the *probability* that each individual sum will give an erroneous value for that bit (or some of bits). For instance, in Example 1, the "sum" $r_1 = u_1 \oplus e_1$ has probability p of giving an erroneous value of u_1 , where p is the crossover probability on the BSC, that is, $\Pr(e_i = 1) = p$ for all i. However, the sum $r_3 \oplus r_5 = u_1 \oplus e_3 \oplus e_5$ has probability 2p(1-p) of giving an erroneous value of u_1 because it gives an erroneous value only when exactly one of e_3 and e_5 is an actual error, that is, has value 1.

APP decoding permits the use of soft-decision demodulation as opposed to the hard-decision demodulation that produces the BSC. In soft-decision demodulation, each received bit r_i is tagged by the demodulator with its probability p_i of being erroneous. With soft-decision demodulation the "sum" $r_1 = u_1 \oplus e_1$ in Example 1 has the probability p_1 of giving an erroneous value of u_1 , whereas the sum $r_3 \oplus r_5 = u_1 \oplus e_3 \oplus e_5$ has probability $p_3(1-p_5) + p_5(1-p_3)$ of giving an erroneous value of u_1 .

It is straightforward to show, cf. Ref. 6, that if $B_1, B_2, \ldots, B_{\delta}$ are the values of δ sums of received bits orthogonal on the information bit u (or on a sum of information bits), then the decision rule for u (or for the sum of information bits) that minimizes error probability from observation of $B_1, B_2, \ldots, B_{\delta}$ is: choose u = 1 (or choose the sum of information bits equal to 1) if and only if

$$\sum_{i=1}^{\delta} w_i B_i > T$$

where the value of B_i is treated as a real number in this sum, where the weighting factor w_i is given by $w_i = 2\log \frac{1-P_i}{P_i}$ wherein P_i is the probability that B_i gives an erroneous value for u, where the threshold T is given by $T = \frac{1}{2} \sum_{i=1}^{\delta} w_i$, and where the information bits are assumed to be independent and equally likely to be 0 or 1. Note that the decision rule for majority decoding can be

1. Note that the decision rule for majority decoding can be written in this form by taking $w_i \equiv 1$. This motivated Massey [6] to introduce the term *threshold decoding* to describe both majority decoding and APP decoding. Majority decoding can be considered as a nonoptimum but simple approximation to APP decoding.

3. THRESHOLD DECODING WITH PARITY CHECKS

It is often convenient to formulate the orthogonal sums discussed above in terms of the parity-check equations of the binary linear code. A *parity check* can be defined as a sum of error bits whose value can be calculated exactly from the received code word, which is equivalent to saying that the sum of the corresponding encoded bits must be zero.
Example 3. Recall that, for the (n = 6, k = 3) code of Example 1, the three sums of received bits orthogonal on the information bit u_1 were $r_1 = u_1 \oplus e_1$, $r_3 \oplus r_5 =$ $u_1 \oplus e_3 \oplus e_5$, and $r_2 \oplus r_6 = u_1 \oplus e_2 \oplus e_6$. If we now add $r_1 = u_1 \oplus e_1$ to each of these sums, we obtain $0 = e_1 \oplus e_1$, $r_1 \oplus r_3 \oplus r_5 = e_1 \oplus e_3 \oplus e_5$, and $r_1 \oplus r_2 \oplus r_6 = e_1 \oplus e_2 \oplus e_6$, which constitute three parity checks orthogonal on the *error bit* e_1 in the sense that each parity check is equal to e_1 plus one or more corrupting error bits, but no error bit corrupts more than one of these parity checks. (Note that e_1 itself corrupts the first of these three parity checks, viz. the trivial parity check 0.) Thus, e_1 will be correctly given by the majority vote of these three parity checks if there is at most one actual error in transmission. Because the first parity check always votes for 0, one can ignore this parity check and say that one decides that e_1 is a 1 if and only if both the second and third parity checks are 1.

One can infer from this example that $\delta - 1$ nontrivial parity checks orthogonal on an error bit (or a sum of error bits) are equivalent to δ sums of received bits orthogonal on an information bit (or a sum of information bits). Moreover, if $A_1, A_2, \ldots, A_{\delta-1}$ are the values of $\delta - 1$ nontrivial parity checks orthogonal on the error bit e (or on a sum of error bits), then the APP decoding rule for e (or for the sum of error bits) from observation of $A_1, A_2, \ldots, A_{\delta-1}$ becomes: decide e = 1 (or decide that the sum of error bits is equal to 1) if and only if

$$\sum_{i=1}^{\delta-1} w_i A_i > T$$

where the value of A_i is treated as a real number in this sum, where the weighting factor w_i is given by $w_i = 2 \log \frac{1-P_i}{P_i}$ wherein P_i is the probability that A_i gives an erroneous value for u, where the threshold T is given by $T = \frac{1}{2} \sum_{i=0}^{\delta-1} w_i$, and where $w_0 = 2 \log \frac{\Pr(e=0)}{\Pr(e=1)}$. The decision rule for majority decoding is obtained by taking $w_i \equiv 1$. For $\delta - 1 = 2$ as in Example 3, the majority decoding rule is decide e = 1 if and only if $A_1 + A_2 > 3/2$ or, equivalently, if and only if $A_1 = A_2 = 1$.

A (reduced) parity-check matrix for an (n, k) linear code is any $(n - k) \times n$ matrix **H** such that $\mathbf{v} = [v_1 \ v_2 \ \cdots \ v_n]$ is a code word if and only if $\mathbf{vH}^T = \mathbf{0}$ where the superscript "*T*" denotes transpose. Again let $\mathbf{r} = \mathbf{v} \oplus \mathbf{e}$ where **r** is the binary received word and **e** is the binary error pattern. Then the syndrome **s** of **r** relative to the paritycheck matrix **H** is defined as $\mathbf{s} = \mathbf{rH}^T$. It follows that $\mathbf{s} = (\mathbf{v} \oplus \mathbf{e})\mathbf{H}^T = \mathbf{eH}^T$ and hence that the syndrome bits are parity checks. In fact, every parity check is either a syndrome bit or a sum of syndrome bits.

If the systematic generator matrix $\mathbf{G} = [\mathbf{I}_k : \mathbf{P}]$ is used for encoding, then the information bits $[u_1 \ u_2 \ \cdots \ u_k] =$ $[v_1 \ v_2 \ \cdots \ v_k]$ determine the "parity bits" $[v_{k+1} \ v_{k+2} \ \cdots \ v_n]$ of the code word as $[v_{k+1} \ v_{k+2} \ \cdots \ v_n] = [v_1 \ v_2 \ \cdots \ v_k] \mathbf{P}$. Thus, \mathbf{v} is a code word if and only if $[v_{k+1} \ v_{k+2} \ \cdots \ v_n] \oplus$ $[v_1 \ v_2 \ \cdots \ v_k] \mathbf{P} = 0$ or, equivalently, if and only if $\mathbf{v}[\mathbf{P}^T \ \vdots \ \mathbf{I}_{n-k}] = \mathbf{0}$. This shows that $\mathbf{H} = [\mathbf{P}^T \ \vdots \ \mathbf{I}_{n-k}]$ is a (reduced) parity-check matrix that we will call the systematic parity-check matrix of the code. Relative to this parity-check matrix, the syndrome becomes $\mathbf{s} = [r_{k+1} \ r_{k+2} \ \cdots \ r_n] \oplus [r_1 \ r_2 \ \cdots \ r_k] \mathbf{P}$. This shows the very useful fact that the syndrome relative to the systematic parity-check matrix can be formed by adding the received parity digits to the parity digits computed from the received information bits.

Example 4. For the (n = 6, k = 3) code of Example 1, the systematic parity-check matrix is

	0	1	1	1	0	0	
$\mathbf{H} =$	1	0	1	0	1	0	
	1	1	0	0	0	1	

The syndrome bits $[s_1 \ s_2 \ s_3] = \mathbf{rH}^T$ are given by $s_1 = r_2 \oplus r_3 \oplus r_4 = e_2 \oplus e_3 \oplus e_4$, $s_2 = r_1 \oplus r_3 \oplus r_5 = e_1 \oplus e_3 \oplus e_5$, and $s_3 = r_1 \oplus r_2 \oplus r_6 = e_1 \oplus e_2 \oplus e_6$. We see that s_2 and s_3 are the $\delta - 1 = 2$ nontrivial parity checks orthogonal on e_1 that we exploited in Example 4. We also note that s_1 and s_3 are two nontrivial parity checks orthogonal on e_2 , and that s_1 and s_2 are two nontrivial parity checks orthogonal on e_3 .

4. THRESHOLD DECODING OF CONVOLUTIONAL CODES

4.1. Preliminaries

Massey [6] formulated the first threshold decoders for multiple-error-correcting convolutional codes. The simplicity of threshold decoders for convolutional codes has led them to dominate practical applications of threshold decoding. We begin here with a brief discussion of those aspects of convolutional codes that are needed to understand threshold decoders for these codes.

In convolutional coding, the information sequences and encoded sequences are semi-infinite sequences that we will represent as *power series* in *D*. For instance, $U(D) = u_0 \oplus$ $u_1D \oplus u_2D^2 \oplus \cdots$ where u_i is the information bit at "time" *i*. In an (n_o, k_o) convolutional code, there are k_o such information sequences $U_1(D), U_2(D), \ldots, U_{k_o}(D)$, and n_o corresponding encoded sequences $V_1(D), V_2(D), \ldots, V_{n_o}(D)$. The encoded sequences are the result of passing the information sequences through a k_o -input/ n_o -output binary finite-state linear system. An example will clarify matters.

Example 5. Consider the $(n_o = 2, k_o = 1)$ convolutional

code with the systematic encoder $\mathbf{G}(\mathbf{D}) = [1 \stackrel{:}{:} P(D)]$ where $P(D) = 1 \oplus D \oplus D^4 \oplus D^6$. The (single) information sequence $U_1(D) = U(D)$ yields the code word

[U(D) : U(D)P(D)], whose two encoded sequences are multiplexed together for transmission over a single channel. For simplicity of notation, let $V(D) = v_0 \oplus v_1 D \oplus v_2 D^2 \oplus \cdots$ denote the sequence of "parity bits" produced by the systematic encoder. Because $V(D) = U(D)P(D) = U(D)(1 \oplus D \oplus D^4 \oplus D^6)$, we see that $v_i = u_i \oplus u_{i-1} \oplus u_{i-4} \oplus u_{i-6}$ for all $i \ge 0$ where it is understood that $u_j = 0$ if j < 0. Because the syndrome relative to the systematic paritycheck matrix can be formed by adding the received parity digits to the parity digits computed from the received



Figure 1. A majority decoder for the (2, 1) convolutional code of Example 5.

information bits, it follows that the syndrome sequence $S(D) = s_0 \oplus s_1 D \oplus s_2 D^2 \oplus \cdots$ can be formed by the simple "linear filter" with a memory of 6 bits shown in Fig. 1. Letting $E(D) = e_0 \oplus e_1 D \oplus e_2 D^2 \oplus \cdots$ and $\Xi(D) =$ $\xi_0 \oplus \xi_1 D \oplus \xi_2 D^2 \oplus \cdots$ denote the error sequences in the information sequence and in the parity-digit sequence, respectively, we further see that the syndrome bits are given by $s_i = e_i \oplus e_{i-1} \oplus e_{i-4} \oplus e_{i-6} \oplus \xi_i$. In particular, we see that $s_6 = e_6 \oplus e_5 \oplus e_2 \oplus e_0 \oplus \xi_6$, $s_4 = e_4 \oplus e_3 \oplus e_0 \oplus \xi_4$, $s_1 = e_1 \oplus e_0 \oplus \xi_1$, and $s_0 = e_0 \oplus \xi_0$ are a set of $\delta - 1 = 4$ parity checks orthogonal on e_0 . Thus, if there are two or fewer actual errors among the 11 error bits that enter into these parity checks, e_0 will be correctly given by majority decoding with threshold T = 5/2, i.e., $e_0^{\Delta} = 1$ if and only if three or more of these parity checks have value 1. One says that the effective constraint length of the convolutional code is $n_E = 11$ bits. We can then feed e_0^{Δ} back to remove e_0 from s_6 , s_4 , and s_1 , following which s_7 , s_5 , s_2 and $s_1{}^{\scriptscriptstyle \Delta} = s_1 \oplus e_0{}^{\scriptscriptstyle \Delta}$ become a set of 4 parity checks orthogonal on e_1 (on the assumption that $e_0^{\Delta} = e_0$). Figure 1 shows the complete double-error correcting majority decoder for the code of this example.

4.2. Early Applications

Codex Corporation was founded in Cambridge, MA, in 1962 as the first organization dedicated solely to the practical application of information-theoretic research. The two innovations that the fledgling company hoped to exploit were Massey's threshold decoders for convolutional codes and Gallager's low-density parity-check (LDPC) codes, the latter a product of a 1960 M.I.T. doctoral thesis and the subject of a 1963 monograph [7]. LDPC codes were "ahead of their day" and defied realization with the discrete logic available in the 1960s. But LDPC codes with Gallager's iterative decoding algorithm have become a hot topic in recent years — reliable transmission at rates extremely close to the capacity of a Gaussian channel has been achieved. Threshold decoders for convolutional codes, however, exhibit a simplicity (cf. Fig. 1) that was virtually ideally suited to realization with the discrete logic available in the 1960s. For this reason, threshold decoders became the first product of Codex Corporation and remained their mainstay product for many years—even though Massey [6] had shown that capacity could not be closely approached with such decoders. Simplicity trumped asymptotics in the 1960s.

It was soon realized at Codex Corporation that most potential applications for threshold decoders were for "bursty channels" in which the actual errors tend to cluster, as opposed to the BSC where the actual errors are scattered. Again convolutional codes were well suited to this problem. Kohlenberg and Forney [8] found by increasing properly the degrees of the nonzero terms in the encoding polynomial P(D) (in the notation of Example 5), a majority decoder acting on the resulting orthogonal parity-checks corrected not only all patterns of $(\delta - 1)/2$ or fewer actual errors among the error bits appearing in these parity checks, but also corrected all bursts of some large length or less. These so-called "diffuse" threshold decoders were in fact the principal coding product of Codex Corporation in the 1960s.

4.3. Self-Orthogonal Codes

A convolutional self-orthogonal code (CSOC) was defined by Massey [6] as a convolutional code with the property that, when the systematic parity-check former is used, all the syndrome bits that check a time-0 error in an information bit are orthogonal on that error bit. Macy [9] and Hagelbarger [10] independently developed majority decoding procedures for CSOCs.

Macy [9] and Robinson et al. [11] independently made the important connection between CSOCs and difference sets. Let $\{i_0, i_i, \ldots, i_q\}$ be a set of q + 1 nonnegative integers. We will say that $\{i_0, i_i, \ldots, i_q\}$ is a *distinctdifferences set* if the (q + 1)q differences of ordered pairs of integers in the set are all distinct. For example, $\{0, 1, 3\}$ is a distinct-differences set because the $(q + 1)q = 3 \cdot 2 = 6$ differences 3 - 0 = 3, 3 - 1 = 2, 1 - 0 = 1, 0 - 3 = -3, 1 - 3 = -2, and 0 - 1 = -1, are all distinct. A distinctdifferences set $\{i_0, i_i, \ldots, i_q\}$ is said to be a *perfect difference* set (or a planar difference set) if its (q + 1)q differences are all distinct and nonzero when taken modulo (q + 1)q + 1. For example, {0, 1, 3} is a perfect difference set because its 6 differences taken modulo 7 are 3, 2, 1, 4, 5, and 6. The distinct-differences set {0, 2, 6} is also a perfect difference set since its 6 differences are 6, 4, 2, -6, -4 and -2, which taken modulo 7 give 6, 4, 2, 1, 3, and 5. However, the distinct-differences set {0, 1, 4} is not a perfect difference set because 4 - 1 = 3 and 0 - 4 = -4 are both 3 when taken modulo 7. Perfect difference sets of q + 1 integers are known to exist whenever q is a power of a prime [12] and perhaps only then.

The convolutional code with encoding polynomial $P(D) = 1 \oplus D \oplus D^4 \oplus D^6$ in Example 5 is a CSOC. The set $\{0, 1, 4, 6\}$ of powers of D appearing in this polynomial is a *perfect difference set*, that is, the $(q + 1)q = 4 \cdot 3 = 12$ differences between ordered pairs of numbers in this set are all distinct and nonzero modulo (q + 1)q + 1 = 13.

The general result embodied in Example 5 is that an $(n_o = 2, k_o = 1)$ convolutional code with systematic

encoder $\mathbf{G}(\mathbf{D}) = [1; P(D)]$ is a CSOC just when the set of powers of *D* appearing in P(D) is a distinct-differences set. One usually desires that the degree of P(D), which is the decoding delay of the threshold decoder and determines the number of delay elements therein, be as small as possible. We point out that the set $\{0, 2, 4, 10\}$ is also a perfect difference set but would be a poorer choice for the degrees of the terms in P(D) than $\{0, 1, 4, 6\}$ because it would give a decoding delay of 10 rather than 6 for the same error-correcting capability. The decoding delay of a CSOC is generally minimized when the distinct-differences set is a perfect difference set, but it takes some skill to find the best perfect difference set. We have restricted our discussion here to $(n_o = 2, k_o = 1)$ CSOCs, but distinctdifferences sets and perfect difference sets also can be used to describe and construct (n_o, k_o) CSOCs with $k_o > 1$ and/or with $n_o > 2$, cf. Ref. 11.

The decoder of Fig. 1 is a so-called *feedback decoder* in which decoded information bits are fed back to remove their effect from the syndrome bits used in future decoding decisions. A decoder without such feedback is called a definite decoder [11] for the convolutional code. Hagelbarger [10] and Robinson et al. [11] independently observed that the feedback could be removed from the majority decoder for a CSOC without reducing the number of errors guaranteed correctable but at the expense of enlarging the effective constraint length n_E . The reason for this is that all the syndrome bits that check each information error bit in a CSOC remain orthogonal even without the removal of past decoded error bits. Recall that in Example 5 the syndrome bits are given by $s_i = e_i \oplus e_{i-1} \oplus e_{i-4} \oplus e_{i-6} \oplus \xi_i$ for all $i \ge 0$. Thus, the syndrome bits that check e_i are $s_i = e_i \oplus e_{i-1} \oplus e_{i-4} \oplus e_{i-6} \oplus \xi_i$, $s_{i+1} = e_{i+1} \oplus e_i \oplus e_{i-3} \oplus e_{i$ $e_{i-5} \oplus \xi_{i+1}, \ s_{i+4} = e_{i+4} \oplus e_{i+3} \oplus e_i \oplus e_{i-2} \oplus \xi_{i+4}, \ \text{and} \ s_{i+6} =$ $e_{i+6} \oplus e_{i+5} \oplus e_{i+2} \oplus e_i \oplus \xi_{i+6}$, which we see are $\delta - 1 = 4$ parity checks orthogonal on e_i with effective constraint length $n_E = 17$. The decoder of Fig. 1 is converted to a definite decoder simply by removing the feedback of e_i^{Δ} to the three modulo-two adders in the "syndrome register." The use of a definite decoder eliminates entirely the error

propagation that results when incorrect decisions are fed back in a feedback decoder. However, the more important fact that this error propagation is very slight for CSOCs was shown in Ref. 11. In virtually all applications of threshold decoding (not only for CSOCs), the performance of the feedback decoder is substantially better than that of the definite decoder.

The (dimensionless) rate R of a (n_o, k_o) convolutional code is defined as $R = k_o/n_o$. Note that $0 < R \le 1$. The bandwidth expansion of the code is 1/R so that highrate codes are preferred in applications where bandwidth is restricted. Wu [13] made an extensive search for good high-rate CSOCs for use in satellite systems. His codes were extensively used in COMSAT and INTELSAT single-channel-per-carrier systems in the 1970s and 1980s in what was perhaps the most significant practical application of threshold decoding yet.

5. THRESHOLD DECODING OF BLOCK CODES

The development of threshold decoding techniques for convolutional codes led to parallel developments for block codes, but without the practical applications for which the convolutional coding systems were so well suited in the 1960s and 1970s. We describe some of these theoretical developments here.

A quasi-cyclic code is a $(n = Mk_o, k = Mn_o)$ linear code having generator matrices and parity-check matrices that can be partitioned into $M \times M$ blocks, each of which is a *circulant matrix*, that is, a square matrix each of whose rows after the first is the right cyclic shift of the previous row. The (6, 3) binary code of Example 4 is a quasi-cyclic code with M = 3. A block self-orthogonal code (BSOC) can be defined as a binary linear code such that, when the systematic parity-check matrix is used, all the syndrome bits that check an error in an information bit are orthogonal on that error bit. The (6, 3) binary code of Example 4 is a BSOC. Townsend and Weldon [14] showed that difference sets play essentially the same role in determining cyclic BSOCs as they do in determining

CSOCs. In the systematic parity-check matrix $\mathbf{H} = [\mathbf{P}^T : \mathbf{I}_3]$ of Example 4, the first row [0 1 1] of the circulant matrix \mathbf{P}^T corresponds to the polynomial $D + D^2$ having {1, 2} as the set of powers of D appearing therein. But {1, 2} is a distinct-differences set whose 2 differences 2 - 1 = 1 and 1 - 2 = -1 are distinct modulo M = 3. This is an illustration of the fact [14] that a binary (n = 2M, k = M) quasi-cyclic code is a BSOC if and only if, in its systematic

parity-check matrix $\mathbf{H} = [\mathbf{P}^T \vdots \mathbf{I}_M]$, the set of powers of D appearing in the polynomial specified by the first row of the circulant matrix \mathbf{P}^T is a distinct-differences set whose differences are also distinct when taken modulo M. Such a code can be completely orthogonalized. If there are q + 1 powers of D in the distinct-differences set, then $\delta = d_{\min} = q + 2$. Perfect difference sets with (q + 1)q + 1 = M are an obvious source for constructing good BSOCs of this type. For instance, using the perfect difference set $\{0, 1, 4, 6\}$ with M = 13 gives a (26, 13) BSOC with $\delta = d_{\min} = 5$. We have restricted our discussion

to quasi-cyclic BSOCs with $k_o = 1$ and $n_o = 2$, but distinctdifferences sets and perfect difference sets also can be used to construct quasi-cyclic BSOCs with $k_o > 1$ and/or with $n_o > 2$, cf. Ref. 14.

Weldon [15] in 1967 showed perfect difference sets can also be used to design majority-decodable cyclic codes as we now explain. First we remark that the so-called *paritycheck polynomial* $h(X) = X^k \oplus h_{k-1}X^{k-1} \dots \oplus h_1X \oplus 1$ of a (n, k) binary cyclic code, which divides the polynomial $X^n \oplus 1$, determines all the parity checks of the cyclic code in the manner that the binary *n*-tuple $[b_{n-1} \ b_{n-2} \dots \ b_1 \ b_0]$ corresponds to (the coefficients of $e_1, \ e_2, \ \dots, \ e_n$ in) a parity check if and only if the polynomial b(X) = $b_{n-1}X^{n-1} \oplus b_{n-2}X^{n-2} \oplus \dots \oplus b_1X \oplus b_0$ is divisible by h(X). For simplicity, we will refer to $[b_{n-1} \ b_{n-2} \dots \ b_1 \ b_0]$ itself as a parity check. Because the code is cyclic, every cyclic shift of a parity check is again a parity check.

Example 6. The set $\{0, 2, 3\}$ of q + 1 = 3 integers is a perfect difference set because the (q + 1)q = 6 differences are nonzero and distinct modulo n = (q + 1)q + 1 = 7. Set b(X) equal to the polynomial whose powers of X correspond to this perfect difference set, that is, $b(X) = X^3 \oplus X^2 \oplus 1$. Now find the polynomial h(X) of largest degree k that divides both b(x) and $X^n \oplus 1$, that is, find the greatest common divisor of b(x) and $X^n \oplus 1$. In this example, b(x) itself divides $X^n \oplus 1 = X^7 \oplus 1$ so that h(X) = b(X) = $X^3 \oplus X^2 \oplus 1$. This h(X) is the parity check polynomial of a cyclic (7, 3) code in which the 7-tuple $[0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1]$ corresponding to b(X) is a parity check. This 7-tuple and its left cyclic shifts by 1 and 3 positions, respectively, viz. $[0\ 0\ 1\ 1\ 0\ 1\ 0]$ and $[1\ 1\ 0\ 1\ 0\ 0]$, constitute a set of $\delta - 1 = 3$ parity checks orthogonal on e_4 . Because the code is cyclic, a set of $\delta - 1 = 3$ parity checks orthogonal on every error bit can be formed by cyclic shifting of these 7-tuples.

Cyclic codes constructed as in Example 6 are called *difference-set cyclic codes* [15] and are always completely orthogonalizable. The particular code in Example 6 happens also to be a BSOC because h(X) = b(x). It is also a maximal-length code; indeed, all maximal-length codes are difference-set cyclic codes and BSOCs. The (21,12) and (73, 45) codes mentioned in Section 2 are also difference-set cyclic codes.

The connection between certain threshold-decodable codes and *finite geometries*, both finite Euclidean geometries and finite projective geometries, was first noted by Rudolph in 1967 [16]. The ramifications of this connection have been of great importance in connection with the general theory of block codes, but of less importance in the practical application of threshold decoding because the finite-geometry codes with parameters suitable for practical implementation were mostly already known. There is a close relationship between perfect difference sets and finite geometries. In fact, Singer's proof [12] that perfect difference sets of q + 1 integers exist whenever q is a power of a prime is based on properties of finite projective geometries.

The codes obtained from finite geometries all rely on viewing the parity checks of the code as *incidence vectors* of some type, for example, as the incidence vectors for points on the lines of the geometry, in which case error bits are regarded as points and parity checks are regarded as lines of the geometry. Perhaps the key contribution was made by Kasami et al. [17] who showed that the Reed-Muller codes shortened by one bit were cyclic finite-geometry codes and subcodes of BCH codes.

6. RECENT DEVELOPMENTS

There has been a steady trickle of new results on threshold decoding since its heyday in the 1960s and 1970s. Threshold decoders continue to be employed in many applications of transmission or storage of information where simplicity and/or high speed in the decoder is a prerequisite. We close this article by mentioning two recent developments that suggest some promising new directions for threshold decoding.

Riedel and Svirid [18] investigated the use of "parallel concatenated" CSOCs in a "turbo-like" structure. They modified the APP decoding algorithm to obtain a soft-in soft-out APP decoder. Iterative decoding of the CSOCs with this algorithm yielded surprisingly good results for the Gaussian channel and for the Rayleigh fading channel.

Meier and Staffelbach [19] devised a cryptanalytic attack on certain stream ciphers that is equivalent to decoding very long maximal-length codes, which we recall from Section 5 are BSOCs, transmitted over a BSC whose crossover probability p is only slightly less than 1/2. Their attack, that is, their decoding algorithm, uses threshold decoding but applied to only a few of the possible parity checks orthogonal on each bit to be decoded (although their paper, which is written for cryptographers, does not state this). This process is iterated until the decoding becomes stable. This "partial decoding" is necessary because the attack must be made with only a small portion of the received code word available to the decoder. Recall from Section 2 that the code word length is $n = 2^m$ but that there are only k = m information bits so that one needs to decode only *m* consecutive received bits to obtain the entire code word. Meier and Staffelbach showed that their attack finds the underlying code word with high probability for surprisingly large channel crossover probabilities. This suggests that their novel way of iterating the decoding of BSOCs may be a useful alternative in data transmission applications to Riedel and Svirid's [18] iterative decoding of CSOCs [18].

BIOGRAPHIES

William W. Wu is the founder of Advanced Technology Mechanization Company (ATMco). He initiated the Internet Domino Technologies and the Universal Transmission Techniques; NSF has supported both programs. He was a PI winner of a multiyear contract from NASA/Glenn Research Center for loss cell recovery in ATM. From 1992 to 1998 he headed the Consultare Group, as advisors to RAFAEL, Israel; TELESAT, Canada; Stanford Telecom in California, COMSAT; POLSPACE, Poland; Motorola; and Hyundai Electronic Industries, Korea. From 1989 to 1991, he was a director at Stanford Telecom. From 1978

to 1989 he was with INTELSAT's executive organ as chief scientist; senior MTS is Communication Engineering, R. and D. Departments. He also participated in INTELSAT *IV, IVA, V VI,* AND-K. From 1967 to 1978, he was a senior scientist at Advance System Div., COMSAT Corporate Headquarters; and at COMSAT Labs. He authored 2 books on digital satellite communications, and has contributed 2 books, 3 encyclopedia, and 60 technical papers. He was the editor of Satellite Communications and Error Coding, IEEE Transactions On Communications. He was also a professional lecture at George Washington University, Washington, D.C., (1977-1989, and 1999), and has given 39 Invited lectures, seminars, and/or short courses worldwide. He is an IEEE Fellow, was chairman for ITU-T SG-XVIII, and a MIT "Distinguished Alumnus" (1998). He has a Ph.D. from the Johns Hopkins University, Maryland MSEE from MIT, Cambridge, Massachusetts, and a BSEE from Purdue University West Lafayette, Indiana.

James L. Massey served on the faculties of the University of Notre Dame, Indiana (1962–1977), the University of California, Los Angeles (1977–1980), and the Swiss Federal Institute of Technology (ETH), Zürich (1980–1998), where he now hold emeritus status. He currently is an adjunct professor at the University of Lund, Sweden.

He has served the *IEEE Transactions on Information Theory* as editor and as associate editor for *Algebraic Coding* and the *Journal of Cryptology* as an associate editor. He is a past president of the IEEE Information Theory Society and of the International Association for Cryptologic Research. He was a founder of Codex Corporation (later a division of Motorola) and of Cylink Corporation, Santa Clara, California.

His awards include the 1988 Shannon Award of the IEEE Information Theory Society, the 1992 IEEE Alexander Graham Bell Medal, and the 1999 Marconi International Fellowship. He is a fellow of the IEEE, a member of the Swiss Academy of Engineering Sciences and the U.S. National Academy of Engineering, an honorary member of the Hungarian Academy of Science, and a foreign member of the Royal Swedish Academy of Sciences.

BIBLIOGRAPHY

- I. S. Reed, A class of multiple-error-correcting codes and the decoding scheme, *IRE Trans. Inform. Theory* **IT-4**: 38–49 (1954).
- R. B. Yale, Error correcting codes and linear recurring sequences, Rept. 34–77, M.I.T. Lincon Lab., Lexington, MA, 1958.
- N. Zierler, On a variation of the first order Reed-Muller codes, Rept. 95, M.I.T. Lincon Lab., Lexington, MA, 1958.
- 4. M. Mitchell, R. Burton, C. Hackett, and R. Schwartz, Coding and operations research, Rept. on Contract AF19(604)-6183, General Electric, Oklahoma City, OK, 1961.
- 5. E. Prange, The use of coset equivalence in the analysis and decoding of group codes, AFCRC Rept. 59-164, Bedford, MA, 1959.
- J. L. Massey, *Threshold Decoding*, M.I.T. Press, Cambridge, MA, 1963.

- R. G. Gallager, Low-Density Parity-Check Codes, M.I.T. Press, Cambridge, MA, 1963.
- A. Kohlenberg and G. D. Forney, Jr., Convolutional coding for channels with memory, *IEEE Trans. Inform. Theory* **IT-14**: 618–626 (1968).
- 9. J. R. Macy, *Theory of Serial Codes*, Ph.D. dissertation, Stevens Inst. Tech., Hoboken, NJ, 1963.
- D. W. Hagelbarger, Recurrent codes for the binary symmetric channel, Lecture Notes, Summer Conf. Theor. Codes, Univ. Michigan, Ann Arbor, MI, 1962.
- J. P. Robinson and A. J. Bernstein, A class of binary recurrent codes with limited error propagation, *IEEE Trans. Inform. Theory* **IT-13**: 106–113 (1967).
- J. Singer, A theorem in finite projective geometry and some applications to number theory, *Trans. Amer. Math. Soc.* 43: 377-385 (1938).
- W. Wu, New convolutional codes, Part I, Part II and Part III, *IEEE Trans. Commun.* COM-23: 942–956 (1975), COM-24: 19–33 and 946–955 (1976).
- R. L. Townsend and E. J. Weldon, Jr., Self-Orthogonal Quasi-Cyclic Codes, *IEEE Trans. Inform. Theory* **IT-13**: 183–195 (1967).
- E. J. Weldon, Jr., Difference set cyclic codes, *Bell Syst. Tech. J.* 45: 1045–1055 (1966).
- L. D. Rudolph, A class of majority logic decodable codes, *IEEE Trans. Inform. Theory* **IT-13**: 305–307 (1967).
- T. Kasami, S. Lin, and W. W. Peterson, New generalizations of the Reed-Muller codes-Part I: Primitive codes, *IEEE Trans. Inform. Theory* **IT-14**: 189-199 (1968).
- S. Ried and Y. V. Svirid, Iterative ("turbo") decoding of threshold decodable codes, *European Trans. Telecomm.* 6: 527-534 (1995).
- W. Meier and O. Staffelbach, Fast correlation attacks on certain stream ciphers, J. Cryptology 1: 159–176 (1989).

TIME DIVISION MULTIPLE ACCESS (TDMA)

PETER JUNG Gerhard-Mercator-Universität Duisburg Duisburg, Germany

1. INTRODUCTION

Communication, stemming from the Latin word for "common," is a most important desire of humankind. The combination of communication with mobility has accelerated the evolution of society worldwide, particularly during the past decade.

The history of mobile radio communication, however, is still young and dates back to the discovery of electromagnetic waves by the German physicist Heinrich Hertz in the nineteenth century. About 100 years ago, Guglielmo Marconi showed that long-haul wireless communication was technically possible using the radio principle based on Hertz's discovery and anticipating what we know as mobile radio today [1].

2586 TIME DIVISION MULTIPLE ACCESS (TDMA)

With the invention of the cellular principle in the early 1970s by engineers of AT&T Bell Labs, the basis for cellular mobile radio systems with high capacity was set [1]. In the early 1980s, the first commercial and civil mobile communication systems like the AMPS (American Mobile Phone Service), the NMT (Nordic Mobile Telecommunication), and the German C450 were introduced, allowing several hundreds of thousands of subscribers [2].

However, technology could not provide digital signal processing at a reasonable degree. Hence, multiple access had to be FDMA (frequency division multiple access). Although FDMA is indispensable for the planning of mobile radio networks, it has some technological drawbacks that led to high-priced base stations and cell phones [2].

However, during the past 20 years, the technological evolution provided us with unprecedented technological possibilities that help to overcome drawbacks of the early mobile radio systems:

- The development of digital signal processing became more and more mature.
- The integration density of microelectronic circuits increased beyond expected limits, providing increasing processing power in small ICs with low power consumption.

In mobile radio, more freedom of choice of the multiple access scheme could be exploited to invent mobile radio systems with more flexibility, higher capacity, and still lower price than the first generation, which relied on FDMA alone.

An increase in system capacity requires base stations that can handle an increased number of traffic channels. With respect to a reduced implementation complexity of the elaborate radiofrequency design of base stations, it is required to support several traffic channels per carrier. Furthermore, to realize radiofrequency front ends with low complexity in cell phones, forward and reverse links should be separated in time. Therefore, TDMA (time division multiple access) provides these assets and hence became the choice for the second generation of mobile radio. Nonetheless, radio network planning remains an important requirement. Hence, TDMA had to be combined with the well-established FDMA, resulting in a hybrid multiple access scheme that is often termed F/TDMA (frequency divided time division multiple access) [3]. These ideas and their further evolutions are considered in what follows.

This article is structured as follows: In Section 2, multiple access principles and hybrid multiple access schemes, which are feasible in mobile radio, are discussed. Section 3 presents signal and system structures used in TDMA systems. The author gives a brief discussion of important TDMA systems for mobile communication in Section 4. An evolution of TDMA toward the third generation of mobile communications, termed UMTS (universal mobile telecommunication system), is presented in Section 5. Section 6 presents concluding remarks.

2. MULTIPLE ACCESS PRINCIPLES AND HYBRID MULTIPLE ACCESS SCHEMES

In Table 1 [3], an overview of the four classic multiple access principles FDMA, TDMA, CDMA (code division multiple access) and SDMA (space division multiple access) is presented. Besides the basic concepts of these multiple access principles, the cellular aspect and an overall evaluation are offered in Table 1.

TDMA, CDMA, and SDMA have become feasible with the introduction of digital technology. However, CDMA was not well understood in civil communications engineering in the 1980s. Only after considerable research effort undertaken in the past two decades, has CDMA been identified as a superb means for mobile multimedia and is therefore deployed in UMTS [4,5]. With the exception of the well-known ANSI/TIA-95, second-generation mobile radio systems did not use CDMA.

In the context of multiple access, SDMA, which requires still expensive smart antennas, has not yet been deployed. However, it is anticipated that SDMA-type technologies will be introduced in upcoming releases of UMTS. SDMA can be regarded as a natural extension of the other three multiple access principles and thus is not considered separately in this article.

According to Table 1, all multiple access principles have specific advantages and drawbacks. To benefit from their advantages and to alleviate the effect of the drawbacks, a combination of multiple access principles, resulting in hybrid multiple access schemes, is recommended.

Considering FDMA, TDMA, and CDMA, four hybrid multiple access schemes are conceivable [3,4], namely, the aforementioned F/TDMA, F/CDMA (frequency divided code division multiple access), T/CDMA (time divided code division multiple access) and F/T/CDMA (frequency and time divided code division multiple access), cf. Fig. 1.

As already discussed, F/TDMA has been chosen for most second-generation mobile radio systems except ANSI/TIA-95, which deploys F/CDMA. Since T/CDMA does not support radio network planning, it has not yet been taken into account for communication systems. F/T/CDMA has been identified as a viable multiple access scheme for the UTRA TDD (UMTS terrestrial radio access time domain duplex) mode. This mode is also known as TD/CDMA and presents an extension of F/TDMA (cf. Section 5).

In what follows, we assume that TDMA is always combined with FDMA, resulting in F/TDMA for the reasons presented in this section. Since TDMA is the most significant part of this hybrid multiple access scheme, the expression TDMA refers to F/TDMA in the sequel.

3. SIGNAL AND SYSTEM STRUCTURES FOR TDMA

3.1. Physical Layer Subscriber Signal Structures

The physical layer subscriber signals carry the data sequences, which shall be transmitted to the receiver. These data sequences consist of encoded subscriber data, which can be any type of information stemming from higher layers (i.e., layers above the physical layer). These subscriber data could for example, be digitally encoded

		Multiple Access F	rinciple	
	FDMA	TDMA	CDMA	SDMA
General Features Basis	Division of system bandwidth B into N_F directly adjacent, disjoint subscriber frequency bands of width $B_u (B_u \ll B)$	Division of the transmission period into directly adjacent, disjoint TDMA frames of duration T_{fr} comprising of N_Z subscriber time slots of duration T_u $(T_u \ll T_{fr})$	Spectrum spreading by using K_g subscriber specific CDMA codes	Division of the cell space into K_g sectors
Subscriber activity	<i>N_F</i> subscribers are simultaneously and continuously active; each subscriber uses a single subscriber frequency band	N_Z subscribers are consecutively active for a short period; each subscriber uses a single subscriber time slot per TDMA frame	K_g subscribers are simultaneously and continuously active; each subscriber uses a single subscriber specific CDMA code	K_g subscribers are simultaneously and continuously active; each subscriber has its own sector
Differentiation between subscriber signals	In the frequency domain	In the time domain	Based on CDMA codes	Based on direction of arrival at receiving antennas
Separating the subscriber signals	by filtering	by deploying synchronization; guard periods between consecutively transmitted subscriber signals are required	by deploying synchronization, single-user detection (SUD) or multi-user detection (MUD)	by using antenna arrays
Area of deployment	Analog and digital	Digital	Digital	Digital
Advantages	Simple; robust; supports network planning; simple equalization	Frequency diversity; receiver is insensitive to time variation of the mobile radio channel; time diversity; high spectral capacity owing to missing intra cell interference; reduced complexity in radio frequency design for cell phones and base stations possible; allows time domain duplexing (TDD)	Frequency diversity; receiver is insensitive to time variation of the mobile radio channel; simple equalizers; interference diversity; soft degradation; no network planning required; flexibility; reduced complexity in radio frequency design for base stations possible	Simple; reduction of multiple access interference; supports network planning; softer handover and space diversity; renunciation on equalizer possible
Disadvantages	Low flexibility; little frequency diversity; receiver sensitive to time variation of the mobile radio channel; little interference diversity; space diversity is necessary; considerable complexity in radio frequency design for cell phones and base stations	Low flexibility; latencies; equalizer is required due to intersymbol interference; little interference diversity; global synchronization of all subscribers, at least in a cell	Low spectral capacity without multi-user detection	Low flexibility; little frequency diversity; receiver sensitive to time variation of the mobile radio channel; reduces interference diversity; low spectral capacity; high implementation complexity of radio frequency design
<i>Cellular Aspects</i> Typical frequency re-use factor	r > 1 due to intercell interference	r > 1 due to intercell interference	$r \approx 1$	r > 1 due to intercell interference
Evaluation	Required for mobile radio; combination with TDMA and/or CDMA is favorable	Applicable in mobile radio; combination with FDMA is strongly suggested and with CDMA is favorable	Applicable in mobile radio; combination with FDMA is strongly suggested and with TDMA is favorable	Applicable in mobile radio; combination with FDMA is strongly suggested and with TDMA and CDMA, respectively,
				is favorable

Table 1. Comparison of Multiple Access Principles [3]



Figure 1. Hybrid multiple access schemes [3,4].

speech. The physical layer subscriber signals must contain signaling information that is required to set up, maintain, and release the connection between transmitter and receiver [3].

For mobile communication, a time-varying multipath channel with an unknown impulse response must be taken into account. To support coherent data detection, channel estimation must be carried out at least once per subscriber time slot. This channel estimation is based on training sequences, which are part of the aforementioned signaling information and which must therefore be embedded in the physical layer subscriber signals. Furthermore, the physical layer subscriber signals are concluded by guard periods of duration T_g in order to guarantee a reasonable separation between consecutive physical layer subscriber signals [3].

As illustrated in Section 2 and Table 1, TDMA allows a subscriber to be active only for a short time before the next period of activity occurs in the next TDMA frame. A typical duration of a subscriber time slot, T_u , is about 0.5 ms, whereas a TDMA frame consists of several subscriber time slots and has a typical duration, T_{fr} , on the order of 5 ms. Hence, the physical layer subscriber signals have a finite duration of T_u . Such signals are usually termed *bursts*.

Figure 2 shows two commonly used burst types [3]. The first burst type [Fig. 2(a)], uses a preamble, which contains the signaling information, including the aforementioned training sequence. When a preamble is used, the aforementioned channel estimation can take place at the beginning of the signal reception. The channel estimate, which is based on noisy samples, is affected by estimation errors due to noise in the received signal. Owing to these estimation errors, the data detection can be only quasicoherent. The noisy channel estimates are fed into the quasi-coherent data detector, which carries out the data detection based on the sample values obtained after the reception of the preamble. Ideally, this quasi-coherent data detection can be carried out without having to store any sample values.

However, in the case of a low correlation time of the mobile radio channel (i.e., at high mobile velocities), the true channel impulse response varies over the duration T_u of the subscriber signals. The error between the noise channel estimate and the true channel impulse response increases nonlinearly with increasing distance from the preamble. In the case of long bursts, this effect leads to considerable systematic errors resulting in dramatic degradations of the quasi-coherent data detection (i.e., of the bit error ratio at a given E_b/N_0).

In order to alleviate this effect, midambles are used instead of preambles [Fig. 2(b)]. In this case, the data are divided in two parts, usually of equal size and half as long as the data carrying part shown in Fig. 2(a). The signaling information is located between these two parts. Then the effect of the above mentioned systematic errors on the bit error ratio is considerably smaller. However, in order to carry out a quasi-coherent data detection, at least those samples associated with the first part of encoded subscriber data must be stored before the channel estimation can be carried out. Nevertheless, thanks to high integration densities in CMOS technology, memory ICs or









Figure 2. Burst types for TDMA [3]. (**a**) Burst type 1: Signaling information as preamble. (**b**) Burst type 2: Signaling information as midamble.



Figure 3. System structure for TDMA [3].

embedded on-chip memories are available at a reasonably low price, thus alleviating this drawback.

A third possibility, using a postamble, suffers from all the drawbacks of the aforementioned two burst types without having further advantages. To the knowledge of the author, this third possibility has not yet been implemented and is not further considered in this article.

3.2. System Structure

Figure 3 shows the corresponding system structure for the physical layer data path between the signal source and the signal sink (cf., e.g., Ref. 3). The system structure consists of a transmitter, a receiver, and the transmission channel.

The transmitter contains source and channel encoders, an interleaver, a burst builder, a modulator, digital and analog filters, the analog RF/IF transmit front end, and at least one transmit antenna. The receiver, particularly the base station receiver, consists of up to K_a receive antennas, K_a RF/IF receive front ends, K_a ADCs (analog-to-digital converters), an adaptive (quasi-) coherent data detector, a de-interleaver, and channel and source decoders.

Usually, hard decided, decoded information combined with the corresponding soft/reliability information are exchanged between the different receiver stages. In this way, a desirably good system performance can be guaranteed.

The system structure shown in Fig. 3 is the basis for the extension to TD/CDMA used in UMTS (see Section 5). There, the corresponding system structure is discussed.

4. TWO IMPORTANT TDMA SYSTEMS FOR MOBILE COMMUNICATION

4.1. Overview

In Fig. 4, the vision generated by the Wireless Strategic Initiative (www.ist-wsi.org) on the further evolution of mobile radio is summarized. The two most important and most successful TDMA systems are the European GSM (global system for mobile communication) [2] and the American UWC-136 (universal wireless communications) [5].

Both TDMA systems started with circuit switched data transmission. In its second phase, GSM was extended to high-speed circuit switched data (HSCSD) with minimal data rates of approximately 14.4 kbit/s and typical data rates between approximately 50 and 60 kbit/s. The corresponding first version of UWC 136 was termed D-AMPS (digital advanced mobile phone service) or IS-54. Both systems were further developed to incorporate packet switching based on GPRS (general packet radio service) and higher data rates using EDGE (enhanced data rates for GSM evolution). Typically, the different EDGE variants provide data rates of approximately 144 kbit/s, with a maximum about 400 kbit/s. GPRS and EDGE evolutions will be part of the family of third-generation mobile communication systems, briefly termed 3G mobile radio systems, with data rates above 400 kbit/s [5-9].

4.2. Global System for Mobile Communication (GSM)

Undoubtedly, the most successful mobile communication system is GSM. Today, GSM provides a multitude of both circuit and packet switched services and applications, including Internet access by using WAP (wireless application protocol) or i-mode. Maximal data rates are currently approximately 50 kbit/s. However, an increase up to approximately 400 kbit/s has already been introduced into the GSM standard.

Table 2 presents important system parameters of GSM, and Fig. 5 gives a comparison between the energy density spectra of well-known digital modulation schemes with those of GMSK (Gaussian minimum shift keying) and the GMSK main impulse [2,3,5,6].

5. TD/CDMA

As mentioned previously, F/TDMA lends itself to further extensions to 3G mobile radio systems. In the early 1990s,



Figure 4. Evolution of wireless communications (source: www.ist-wsi.org).

the combination of F/TDMA with CDMA was presented, resulting in the hybrid multiple access scheme F/T/CDMA already considered in Section 2 and Fig. 1. Now, up to K subscribers could operate simultaneously within a TDMA time slot [3,5].

A major problem to be solved in a CDMA-based system is the near-far problem. In order to alleviate the necessity of fast power control, which cannot be provided at high velocities when a TDMA component is used, multiuser detection is a must in F/T/CDMA. It has been shown that suboptimal joint detection (JD) techniques based on block linear equalization and block decision feedback equalization lend themselves as viable means for such F/T/CDMA-based systems.

The channel estimation to be used must be capable of estimating a multitude of simultaneous transmission channels in the uplink. Steiner proposed a means of generating good training sequences for this purpose and proposed a novel channel estimator [10], sometimes termed the Steiner estimator. These milestone developments—the JD techniques and the Steiner estimator—paved the way toward what is known as TD/CDMA or UTRA TDD, today [3,5].

By moderately modifying the system structure shown in Fig. 3, the corresponding system structure for TD/CDMA can be found (Fig. 6). In Fig. 7, the physical layer subscriber signal is shown schematically. Both system and signal structures are consequent evolutions toward more multimedia in mobile communications, and it can be anticipated that TDMA components will remain important in the future.

6. CONCLUSIONS

TDMA as a viable means to solve the multiple access problem in mobile communication was discussed in this article. Besides giving an overview of the four classic multiple access principles, hybrid multiple access schemes were illustrated. Furthermore, we discussed both physical layer signal structures and the corresponding system structures for TDMA systems deployed in mobile communications. Finally, important TDMA systems and the evolution toward 3G mobile radio systems were briefly sketched.

To conclude, it should be mentioned that TDMA also provides excellent features in short-range communication systems. Therefore, wireless local area networks and some future wireless communication system concepts "beyond 3G" rely on TDMA components.

BIOGRAPHY

Peter Jung received the diploma (M.Sc. equiv.) in physics from the University of Kaiserslautern, Germany, in 1990, and the Dr.-Ing. (Ph.D.EE equiv.) and Dr.-Ing. habil. (D.Sc.EE equiv.), both in electrical engineering with a focus on microelectronics and communications technology, from the University of Kaiserslautern in 1993 and 1996, respectively. In 1996, he became private educator (equiv. to reader) at the University of Kaiserslautern and in 1998 also at Technical University of Dresden, Germany. From March 1998 to May 2000, he was with

Multiple Access Scheme		F/TDMA				
Modulation scheme	Phases 1,2,2+:	GMSK (Gaussian minimum shift keying)				
	EDGE:	GMSK, (3/8) π -Offset-8-PSK (8-ary Phase Shift Keying) with spectral forming by GMSK main impulse				
Subscriber bandwidth	200 kHz					
Symbol rate	270.833 ksymbols	/s				
Duration of a TDMA frame, T_{fr} 4.615 ms						
Number of subscriber time slots 8 per TDMA frame						
Uplink (reverse link) frequency 880915 (i) bands 17201785 (i) 19301990 (i) (i)		(GSM 900, e.g., German D networks) (DCS 1800, e.g., German E networks) (American GSM 1900)				
Downlink (forward link) frequency bands	935960 18051880 18501910	(GSM 900, e.g., German D networks) (DCS 1800, e.g., German E networks) (American GSM 1900)				
Maximal information rate per subscriber	Speech full rate: half rate: enhanced full rate Data TCH/9.6 (ph phase 2+ HSCSD phase 2+ GPRS:	13 kbit/s 6.5 kbit/s 8: 12.2 kbit/s ase 1): 9.6 kbit/s 115.2 kbit/s 171.2 kbit/s (four coding schemes, today, only coding scheme CS-2 is used; three classes of mobile equipment; 18 multi-slot classes, today, classes 4 and 8 are usually implemented, cf. www.csdmag.com)				
	EDGE:	packet switching with data rates of minimally 384 kbit/s for velocities below 100 km/h; packet switching with data rates of minimally 144.4 kbit/s for velocities between 100 km/h and 250 km/h				
Frequency hopping (optional)	1 hop per TDMA f	rame, i.e., 217 hops/s				

Table 2. Important System Parameters of GSM [5,6]



Figure 5. Energy density spectra.

Siemens AG, Bereich Halbleiter, now Infineon Technologies, as Director of Cellular Innovation and later Senior Director of Concept Engineering Wireless Baseband. In June 2000, he became full professor and chair for communication technology at Gerhard-Mercator-University Duisburg and director of the Fraunhofer-Institut für Mikroelekronische Schaltungen und Systeme (IMS), Duisburg. In 1995, he was co-recipient of the best paper award at the ITG-Fachtagung Mobile Kommunikation, Ulm, Germany, and in 1997, he was co-recipient of the Johann-Philipp-Reis-Award for his work on multicarrier CDMA mobile radio systems. His areas of interest include wireless communication technology, softwaredefined radio, and system-on-a-chip integration of communication systems.

BIBLIOGRAPHY

- G. Calhoun, Digital Cellular Radio, Norwood, Mass., Artech House, 1988.
- 2. M. Mouly and M.-B. Pautet, The GSM System for Mobile Communications, 1992.
- 3. P. Jung, Analyse und Entwurf digitaler Mobilfunksysteme, Stuttgart, Teubner, 1997.
- P. W. Baier, P. Jung, A. Klein, Taking the challenge of multiple access for third generation mobile radio systems — a European view, *IEEE Commun. Mag.* 82–89 (Feb. 1996).
- T. Ojanperä and R. Prasad, eds., Wideband CDMA for Third Generation Mobile Communications, Boston, Artech House, 1998.

2592 TIME DIVISION MULTIPLE ACCESS (TDMA)



Figure 6. System structure for TD/CDMA [3].



Figure 7. Data structure in a TD/ CDMA burst [3].

- A. Furuskär, S. Mazur, F. Müller, and H. Olofsson, EDGE: Enhanced Data Rates for GSM and TDMA/136 Evolution, *IEEE Personal Commun.* 6: 56-66 (1999).
- R. Prasad, W. Mohr, and W. Konhäuser, eds., *Third Genera*tion Mobile Communication Systems, Boston, Artech House, 2000.
- 8. B. Walke, M. P. Althoff, and P. Seidenberg, UMTS-Ein Kurs, Weil der Stadt, Schlembach, 2001.
- 9. H. Holma and A. Toskala, eds., WCDMA for UMTS, Chichester, Wiley, 2000.
- B. Steiner and P. Jung, Optimum and suboptimum channel estimation for the uplink of CDMA mobile radio systems with joint detection, *European Transactions on Telecommunications and Related Technologies* (*ETT*) 5: 39-50 (1994).

TRANSFORM CODING

VIVEK GOYAL Digital Fountain Inc. Fremont, California

1. INTRODUCTION

Transform coding is a type of source coding characterized by a modular design that includes a linear transformation of the original data and scalar quantization of the resulting coefficients. It arises from applying the "divide and conquer" principle to lossy source coding. This principle of breaking a major problem into smaller problems that can be more easily understood and solved is central in engineering and computational science. The resulting modular design is advantageous for implementation, testing, and component reuse.

Everyday compression problems are unmanageable without a divide-and-conquer approach. Effective compression of images, for example, depends on the tendencies of pixels to be similar to their neighbors, or to differ in partially predictable ways. These tendencies, arising from the continuity, texturing, and boundaries of objects, the similarity of objects in an image, gradual lighting changes, an artist's technique and color palette, or similar may extend over an entire image with a quarter-million pixels. Yet the most general way to utilize the probable relationships between pixels (later described as *unconstrained source coding*) is infeasible for this many pixels. In fact, 16 pixels is a lot for an unconstrained source code.

To conquer the compression problem — allowing, for example, more than 16 pixels to be encoded simultaneously — state-of-the-art lossy compressors divide the encoding operation into a sequence of three relatively simple steps: the computation of a linear transformation of the data designed primarily to produce uncorrelated coefficients, separate quantization of each scalar coefficient, and entropy coding. This process is called *transform coding*. In image compression, a square image with N pixels is typically processed with simple linear transforms (often discrete wavelet transforms) of size $N^{1/2}$.

This article explains the fundamental principles of transform coding with reference to abstract sources. These principles apply equally well to images, audio, video, and various other types of data.

1.1. Source Coding

Source coding is to represent information in bits, with the natural aim of using a small number of bits. When the information can be exactly recovered from the bits, the source coding or *compression* is called *lossless*; otherwise, it is called *lossy*. The transform codes in this article are lossy. However, lossless entropy codes appear as components of transform codes, so both lossless and lossy compression are of present interest.

In our discussion, the "information" is denoted by a real column vector $x \in \mathbb{R}^N$ or a sequence of such vectors. A vector might be formed from pixel values in an image or by sampling an audio signal; $K \cdot N$ pixels can be arranged as

a sequence of K vectors of length N. The vector length N is defined such that each vector in a sequence is encoded independently. For the purpose of building a mathematical theory, the source vectors are assumed to be realizations of a random vector \mathbf{x} with a known distribution. The distribution could be purely empirical.

A source code is composed of two mappings: an encoder and a decoder. The encoder maps any vector $x \in \mathbb{R}^N$ to a finite string of bits, and the decoder maps any of these strings of bits to an approximation $\hat{x} \in \mathbb{R}^N$. The encoder mapping can always be factored as $\gamma \circ \alpha$, where α is a mapping from \mathbb{R}^N to some discrete set \mathcal{I} and γ is an invertible mapping from \mathcal{I} to strings of bits. The former is called a lossy encoder and the latter a lossless code or an entropy code. The decoder inverts γ and then approximates x from the index $\alpha(x) \in \mathcal{I}$. This is shown in the top half of Fig. 1. It is assumed that communication between the encoder and decoder is perfect.

To assess the quality of a lossy source code, we need numerical measures of approximation accuracy and description length. The measure for description length is simply the expected number of bits output by the encoder divided by N; this is called the *rate* in bits per scalar sample and denoted by R. Here we will measure approximation accuracy by squared Euclidean norm divided by the vector length:

$$d(x, \hat{x}) = rac{1}{N} \|x - \hat{x}\|^2 = rac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

This accuracy measure is conventional and usually leads to the easiest mathematical results, though the theory of source coding has been developed with quite general measures [1]. The expected value of $d(\mathbf{x}, \hat{\mathbf{x}})$ is called the mean-squared error (MSE) *distortion* and is denoted by $D = E[d(\mathbf{x}, \hat{\mathbf{x}})]$. The normalizations by N make it possible to fairly compare source codes with different lengths.

Fixing N, a theoretical concept of optimality is straightforward: A length-N source code is *optimal* if no other length-N source code with at most the same rate has lower distortion. This concept is of dubious value. First, it is very difficult to check the optimality of a source code. Local optimality — being assured that small perturbations of α and β will not improve performance — is often the best that can be attained [14]. Second, and of more practical consequence, a system designer gets to choose the value of N. It can be as large as the total size of the data set — like the number of pixels in an image — but can also be smaller, in which case the data set is interpreted as a sequence of vectors.

There are conflicting motives in choosing N. Compression performance is related to the predictability of one part of x from the rest. Since predictability can only increase from having more data, performance is usually improved by increasing N. (Even if the random variables producing each scalar sample are mutually independent, the optimal performance is improved by increasing N; however, this "packing gain" effect is relatively small [14].) The conflict comes from the fact that the computational complexity of encoding is also increased. This is particularly dramatic if one looks at complexities of optimal source codes. The



Figure 1. Any source code can be decomposed so that the encoder is $\gamma \circ \alpha$ and the decoder is $\beta \circ \gamma^{-1}$, as shown at top. γ is an entropy code and α and β are the encoder and decoder of an *N*-dimensional quantizer. In a transform code, α and β each have a particular constrained structure. In the encoder, α is replaced with a linear transform *T* and a set of *N* scalar quantizer encoders. The intermediate y_i s are called *transform coefficients*. In the decoder, β is replaced with *N* scalar quantizer decoders and another linear transform *U*. Usually $U = T^{-1}$.

obvious way to implement an optimal encoder is to search through the entire codebook, giving running time exponential in N. Other implementations reduce running time while increasing memory usage [24].

State-of-the-art source codes result from an intelligent compromise. There is no attempt to realize an optimal code for a given value of N because encoding complexity would force a small value for N. Rather, source codes that are good, but plainly not optimal, are used. Their lower complexities make much larger N values feasible. This has eloquently been called "the power of imperfection" [5]. The paradoxical conclusion is that the best codes to use in practice are *suboptimal*.

1.2. Constrained Source Coding

Transform codes are the most often used source codes because they are easy to apply at any rate and even with very large values of N. The essence of transform coding is the modularization shown in the bottom half of Fig. 1. The mapping α is implemented in two steps. First, an invertible linear transform of the source vector x is computed, producing y = Tx. Each component of y is called a *transform coefficient*. The N transform coefficients are then quantized independently of each other by Nscalar quantizers. This is called *scalar quantization* since each scalar component of y is treated separately. Finally, the quantizer indices that correspond to the transform coefficients are compressed with an entropy code to produce the sequence of bits that represent the data.

To reconstruct an approximation of x, the decoder essentially reverses the steps of the encoder. The action of the entropy coder can be inverted to recover the quantizer indices. Then the decoders of the scalar quantizers produce a vector \hat{y} of estimates of the transform coefficients. To complete the reconstruction, a linear transform is applied to \hat{y} to produce the approximation \hat{x} . This final step usually uses the transform T^{-1} , but for generality the transform is denoted U.

Most source codes cannot be implemented in the two stages of linear transform and scalar quantization. Thus, a transform code is an example of a *constrained source code*. Constrained source codes are, loosely speaking, source codes that are suboptimal but have low complexity. The simplicity of transform coding allows large values of Nto be practical. Computing the transform T requires at most N^2 multiplications and N(N-1) additions. Specially structured transforms—such as discrete Fourier, cosine, and wavelet transforms—are often used to reduce the complexity of this step, but this is merely icing on the cake. The great reduction from the exponential complexity of a general source code to the (at most) quadratic complexity of a transforms code comes from using linear transforms and scalar quantization.

The difference between constrained and unconstrained source codes is demonstrated by the partition diagrams in Fig. 2. In these diagrams, the cells indicate which source vectors are encoded to the same index and the dots are the reconstructions computed by the decoder. Four locally optimal fixed-rate source codes with 12-element codebooks were constructed. The two-dimensional, jointly Gaussian source is the same as that used later in Fig. 3.

The partition for an unconstrained code shares the symmetries of the source density but is otherwise complicated because the cell shapes are arbitrary. Encoding is difficult because there is no simple way to get around using both components of the source vector simultaneously in computing the index.

Encoding with a transform code is easier because after the linear transform the coefficients are quantized separately. This gives the structured alignment of partition cells and of reconstruction points shown.

It is fair to ask why the transform is linear. In two dimensions, one might imagine quantizing in polar



Figure 2. Partition diagrams for (a) unconstrained code (D = 0.055); (b) transform code (D = 0.066); (c) angular quantization (D = 0.116); (d) polar-coordinate quantization (D = 0.069).



Figure 3. Illustration of various basis changes: (a) a basis change generally includes a nonhypercubic partition; (b) a singular transformation gives a partition with unbounded cells; (c) a Karhunen-Loève transform is an orthogonal transform that aligns the partitioning with the axes of the source PDF. The source is depicted by level curves of the PDF (left). The transform coefficients are separately quantized with uniform quantizers (center). The induced partitioning is then shown in the original coordinates (right).

coordinates. Two examples of partitions obtained with separate quantization of radial and angular components are shown in Figs. 2c and 2d, and these are as elegant as the partition obtained with a linear transform. Yet nonlinear transformations — even transformations to polar coordinates — are rarely used in source coding. With arbitrary transformations, the approximation accuracy of the transform coefficients does not easily relate to the accuracy of the reconstructed vectors. This makes designing quantizers for the transform coefficients more difficult. Also, allowing nonlinear transformations reintroduces the design and encoding complexities of unconstrained source codes.

Constrained source codes need not use transforms to have low complexity. Techniques described in the literature [8,14] include those based on lattices, sorting, and tree-structured searching; none of these techniques is as popular as transform coding.

2. THE STANDARD MODEL AND ITS COMPONENTS

The standard theoretical model for transform coding has the strict modularity shown in the bottom half of Fig. 1, meaning that the transform, quantization, and entropy coding blocks operate independently. In addition, the entropy coder can be decomposed into N parallel entropy coders so that the quantization and entropy coding operate independently on each scalar transform coefficient.

This section briefly describes the fundamentals of entropy coding and quantization to provide background for our later focus on the optimization of the transform. The final part of this section addresses the allocation of bits among the N scalar quantizers. Additional information can be found in the literature [3,8,12,14].

2.1. Entropy Codes

Entropy codes are used for lossless coding of discrete random variables. Consider the discrete random variable \mathbf{z} with alphabet \mathcal{I} . An entropy code γ assigns a unique binary string, called a *codeword*, to each $i \in \mathcal{I}$ (see Fig. 1).

Since the codewords are unique, an entropy code is always invertible. However, we will place more restrictive conditions on entropy codes so they can be used on sequences of realizations of \mathbf{z} . The *extension* of γ maps the finite sequence (z_1, z_2, \ldots, z_k) to the concatenation of the outputs of γ with each input, $\gamma(z_1)\gamma(z_2)\cdots\gamma(z_k)$. A code is called *uniquely decodable* if its extension is one to one. A uniquely decodable code can be applied to message sequences without adding any "punctuation" to show where one codeword ends and the next begins. In a *prefix code*, no codeword is the prefix of any other codeword. Prefix codes are guaranteed to be uniquely decodable.

A trivial code numbers each element of \mathcal{I} with a distinct index in $\{0, 1, \ldots, |\mathcal{I}| - 1\}$ and maps each element to the binary expansion of its index. Such a code requires $\lceil \log_2 |\mathcal{I}| \rceil$ bits per symbol. This is considered the *lack* of an entropy code. The idea in entropy code design is to minimize the mean number of bits used to represent \mathbf{z} at

the expense of making the worst-case performance worse. The expected code length is given by

$$L(\gamma) = E[\ell(\gamma(\mathbf{z}))] = \sum_{i \in \mathcal{I}} p_{\mathbf{z}}(i)\ell(\gamma(i))$$

where $p_{\mathbf{z}}(i)$ is the probability of symbol *i* and $\ell(\gamma(i))$ is the length of $\gamma(i)$. The expected length can be reduced if short codewords are used for the most probable symbols — even if this means that some symbols will have codewords with more than $\lceil \log_2 |\mathcal{I}| \rceil$ bits.

The entropy code γ is called *optimal* if it is a prefix code that minimizes $L(\gamma)$. Huffman codes are examples of optimal codes. The performance of an optimal code is bounded by

$$H(\mathbf{z}) \le L(\gamma) < H(\mathbf{z}) + 1 \tag{1}$$

where

$$H(\mathbf{z}) = -\sum_{i \in \mathcal{I}} p_{\mathbf{z}}(i) \log_2 p_{\mathbf{z}}(i)$$

is the *entropy* of **z**.

The up to one bit gap in Eq. (1) is ignored in the remainder of the article. If $H(\mathbf{z})$ is large, this is justified simply because one bit is small compared to the code length. Otherwise note that $L(\gamma) \approx H(\mathbf{z})$ can attained by coding blocks of symbols together; this is detailed in any information theory or data compression textbook.

2.2. Quantizers

A quantizer q is a mapping from a source alphabet \mathbb{R}^N to a reproduction codebook $\mathcal{C} = \{\hat{x}_i\}_{i \in \mathcal{I}} \subset \mathbb{R}^N$, where \mathcal{I} is an arbitrary countable index set. Quantization can be decomposed into two operations $q = \beta \circ \alpha$, as shown in Fig. 1. The lossy encoder $\alpha : \mathbb{R}^N \to \mathcal{I}$ is specified by a partition of \mathbb{R}^N into partition cells $S_i = \{x \in \mathbb{R}^N \mid \alpha(x) = i\},$ $i \in \mathcal{I}$. The reproduction decoder $\beta : \mathcal{I} \to \mathbb{R}^N$ is specified by the codebook \mathcal{C} . If N = 1, the quantizer is called a scalar quantizer; for N > 1, it is a vector quantizer.

The quality of a quantizer is determined by its distortion and rate. The MSE distortion for quantizing random vector $\mathbf{x} \in \mathbb{R}^{N}$ is

$$D = E[d(\mathbf{x}, q(\mathbf{x}))] = N^{-1}E[\|\mathbf{x} - q(\mathbf{x})\|^2]$$

The rate can be measured in several ways. The lossy encoder output $\alpha(\mathbf{x})$ is a discrete random variable that usually should be entropy coded because the output symbols will have unequal probabilities. Associating an entropy code γ to the quantizer gives a *variable-rate quantizer* specified by (α, β, γ) . The rate of the quantizer is the expected code length of γ divided by N. Not specifying an entropy code (or specifying the use of fixed-rate binary expansion) gives a *fixed-rate quantizer* with rate $R = N^{-1} \log_2 |\mathcal{I}|$. Measuring the rate by the idealized performance of an entropy code gives $R = N^{-1} H(\alpha(\mathbf{x}))$; the quantizer in this case is called *entropy-constrained*.

The optimal performance of variable-rate quantization is at least as good as that of fixed-rate quantization, and entropy-constrained quantization is better yet. However, entropy coding adds complexity, and variable-length output can create difficulties such as buffer overflows. Furthermore, entropy-constrained quantization is only an idealization since an entropy code will generally not meet the lower bound in Eq. (1).

2.2.1. Optimal Quantization. An *optimal quantizer* is one that minimizes the distortion subject to an upper bound on the rate or minimizes the rate subject to an upper bound on the distortion. Because of simple shifting and scaling properties, an optimal quantizer for a scalar **x** can be easily deduced from an optimal quantizer for the normalized random variable $(\mathbf{x} - \mu_{\mathbf{x}})/\sigma_{\mathbf{x}}$, where $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ are the mean and standard deviation of **x**, respectively. One consequence of this is that optimal quantizers have performance

$$D = \sigma^2 g(R) \tag{2}$$

where σ^2 is the variance of the source and g(R) is the performance of optimal quantizers for the normalized source. Equation (2) holds, with a different function g, for any family of quantizers that can be described by its operation on a normalized variable, not just optimal quantizers.

Optimal quantizers are difficult to design, but locally optimal quantizers can be numerically approximated by an iteration in which α , β , and γ are separately optimized, in turn, while keeping the other two fixed. For details on each of these optimizations and the difficulties and properties that arise, see Refs. 5 and 14.

Note that the rate measure affects the optimal encoding rule because $\alpha(x)$ should be the index that minimizes a Lagrangian cost function including both rate and distortion; for example

$$\alpha(x) = \operatorname*{argmin}_{i \in \mathcal{I}} \left[\frac{1}{N} \ell(\gamma(i)) + \lambda \frac{1}{N} \|x - \beta(i)\|^2 \right]$$

is an optimal lossy encoder for variable-rate quantization. (By fixing the relative importance of rate and distortion, the Lagrange multiplier λ determines a rate-distortion operating point among those possible with the given β and γ .) Only for fixed-rate quantization does the optimal encoding rule simplify to finding the index corresponding to the nearest codeword.

In some of the more technical discussions that follow, one property of optimal decoding is relevant: The optimal decoder β computes

$$\beta(i) = E[\mathbf{x} \mid \mathbf{x} \in S_i]$$

which is called *centroid reconstruction*. The conditional mean of the cell, or centroid, is the minimum MSE estimate [26].

2.2.2. High-Resolution Quantization. For most sources, it is impossible to analytically express the performance of optimal quantizers. Thus, aside from using Eq. (2), approximations must suffice. Fortunately, approximations obtained when it is assumed that the quantization is very fine are reasonably accurate even at low to moderate rates. Details on this "high resolution" theory for both scalars and vectors can be found in Refs. 7 and 14 and other sources cited therein.

Let $f_{\mathbf{x}}(x)$ denote the probability density function (PDF) of the scalar random variable \mathbf{x} . High-resolution analysis is based on approximating $f_{\mathbf{x}}(x)$ on the interval S_i by its value at the midpoint. Assuming that $f_{\mathbf{x}}(x)$ is smooth, this approximation is accurate when each S_i is short.

Optimization of scalar quantizers turns into finding the optimal lengths for the S_i s, depending on the PDF $f_x(x)$. One can show that the performance of optimal fixed-rate quantization is approximately

$$D \approx \frac{1}{12} \left(\int_{\mathbb{R}} f_{\mathbf{x}}^{1/3}(x) \, dx \right)^3 2^{-2R} \tag{3}$$

Evaluating this for a Gaussian source with variance σ^2 gives

$$D \approx \frac{1}{2} 3^{1/2} \pi \sigma^2 2^{-2R}$$
 (4)

For entropy-constrained quantization, high-resolution analysis shows that it is optimal for each S_i to have equal length [9]. A quantizer that partitions with equal-length intervals is called *uniform*. The resulting performance is

$$D \approx \frac{1}{12} 2^{2h(\mathbf{x})} 2^{-2R}$$
 (5)

where

$$h(\mathbf{x}) = -\int_{\mathbb{R}} f_{\mathbf{x}}(x) \log_2 f_{\mathbf{x}}(x) \, dx$$

is the *differential entropy* of \mathbf{x} . For Gaussian random variables, Eq. (5) simplifies to

$$D \approx \frac{\pi e}{6} \sigma^2 2^{-2R} \tag{6}$$

Summarizing Eqs. (3)-(6), the lesson from high resolution quantization theory is that quantizer performance is described by

$$D \approx c\sigma^2 2^{-2R} \tag{6}$$

where σ^2 is the variance of the source and *c* is a constant that depends on the normalized density of the source and the type of quantization (fixed-rate, variable-rate or entropy-constrained). This is consistent with Eq. (2).

The computations we have made are for scalar quantization. For vector quantization, the best performance in the limit as the dimension N grows is given by the distortion rate function [1]. For a Gaussian source this bound is $D = \sigma^2 2^{-2R}$. The approximate performance given by Eq. (6) is worse by a factor of only $\pi e/6$ (≈ 1.53 dB). This can be expressed as a redundancy $\frac{1}{2} \log_2(\frac{\pi e}{6}) \approx 0.255$ bits. Furthermore, a numerical study has shown that for a wide range of memoryless sources, the redundancy of entropy-constrained uniform quantization is at most 0.3 bits per sample at all rates [6].

2.3. Bit Allocation

Coding (quantizing and entropy coding) each transform coefficient separately splits the total number of bits among the transform coefficients in some manner. Whether done with conscious effort or implicitly, this is a *bit allocation* among the components. Bit allocation problems can be stated in a single common form: One is given a set of quantizers described by their distortion-rate performances as

$$D_i = g_i(R_i), \quad R_i \in \mathcal{R}_i, \quad i = 1, 2, \dots, N$$

Each set of available rates \mathcal{R}_i is a subset of the nonnegative real numbers and may be either discrete or continuous. The problem is to minimize the average

distortion
$$D = N^{-1} \sum_{i=1}^{N} D_i$$
 given a maximum average rate

$$R = N^{-1} \sum_{i=1}^{N} R_i.$$

As is often the case with optimization problems, bit allocation is easy when the parameters are continuous and the objective functions are smooth. Subject to a few other technical requirements, parametric expressions for the optimal bit allocation can be found elsewhere [22,29]. The techniques used when the \mathcal{R}_i s are discrete are quite different and play no role in forthcoming results [30].

If the average distortion can be reduced by taking bits away from one component and giving them to another, the initial bit allocation is not optimal. Applying this reasoning with infinitesimal changes in the component rates, a necessary condition for an optimal allocation is that the slope of each g_i at R_i is equal to a common constant value. A tutorial treatment of this type of optimization has been published [25].

The approximate performance given by Eq. (7) leads to a particularly easy bit allocation problem with

$$g_i = c_i \sigma_i^2 2^{-2R_i}, \quad \mathcal{R}_i = [0, \infty), \quad i = 1, 2, \dots, N$$
 (8)

Ignoring the fact that each component rate must be nonnegative, an equal-slope argument shows that the optimal bit allocation is

$$R_i = R + rac{1}{2}\log_2rac{c_i}{\left(\prod\limits_{i=1}^N c_i
ight)^{1/N}} + rac{1}{2}\log_2rac{\sigma_i^2}{\left(\prod\limits_{i=1}^N \sigma_i^2
ight)^{1/N}}$$

With these rates, all the D_i s are equal and the average distortion is

$$D = \left(\prod_{i=1}^{N} c_i\right)^{1/N} \left(\prod_{i=1}^{N} \sigma_i^2\right)^{1/N} 2^{-2R}.$$
 (9)

This solution is valid when each R_i given above is nonnegative. For lower rates, the components with smallest $c_i \cdot \sigma_i^2$ are allocated no bits and the remaining components have correspondingly higher allocations.

2.3.1. Bit Allocation With Uniform Quantizers. With uniform quantizers, bit allocation is nothing more than choosing a step size for each of the N components. The equal-distortion property of the analytical bit allocation solution gives a simple rule: Make all of the step sizes equal. This will be referred to as "lazy" bit allocation.

Our development indicates that lazy allocation is optimal when the rate is high. In addition, numerical

studies have shown that lazy allocation is nearly optimal as long as the minimal allocated rate is at least 1 bit [12,13].

3. OPTIMAL TRANSFORMS

It has taken some time to set the stage, but we are now ready for the main event of designing the analysis transform T and the synthesis transform U. Throughout this section the source \mathbf{x} is assumed to have mean zero, and $R_{\mathbf{x}}$ denotes the covariance matrix $E[\mathbf{x}\mathbf{x}^T]$, where T denotes the transpose. The source is often—but not always—jointly Gaussian.

A signal given as a vector in \mathbb{R}^N is implicitly represented as a series with respect to the standard basis. An invertible analysis transform *T* changes the basis. A change of basis does not alter the information in a signal, so how can it affect coding efficiency? Indeed, if arbitrary source coding is allowed after the transform, it does not. The motivating principle of transform coding is that *simple* coding may be more effective in the transform domain than in the original signal space. In the standard model, "simple coding" corresponds to the use of scalar quantization and scalar entropy coding.

3.1. Visualizing Transforms

Beyond two or three dimensions, it is difficult to visualize vectors — let alone the action of a transform on vectors. Fortunately, most people already have an idea of what a linear transform does: it combines rotating, scaling, and shearing such that a hypercube is always mapped to a parallelepiped.

In two dimensions, the level curves of a zero-mean Gaussian density are ellipses centered at the origin with collinear major axes, as shown in the left panels of Fig. 3. The middle panel of Fig. 3a shows the level curves of the joint density of the transform coefficients after a more or less arbitrary invertible linear transformation. A linear transformation of an ellipse is still an ellipse, although its eccentricity and orientation (direction of major axis) may have changed.

The grid in the middle panel indicates the cell boundaries in uniform scalar quantization, with equal step sizes, of the transform coefficients. The effect of inverting the transform is shown in the right panel; the source density is returned to its original form and the quantization partition is linearly deformed. The partition in the original coordinates, as shown in the right panel, is what is truly relevant. It shows which source vectors are mapped to the same symbol, thus giving some indication of the average distortion. Looking at the number of cells with appreciable probability gives some indication of the rate.

A singular transform is a degenerate case. As shown in the middle panel of Fig. 3b, the transform coefficients have probability mass only along a line. (A *line segment* is an ellipse with unit eccentricity.) Inverting the transform is not possible, but we may still return to the original coordinates to view the partition induced by quantizing the transform coefficients. The cells are unbounded in one direction, as shown in the right panel. This is undesirable unless variation of the source in the direction in which the cells are unbounded is very small.

3.1.1. Shapes of Partition Cells. Although better than unbounded cells, the parallelogram-shaped partition cells that arise from arbitrary invertible transforms are inherently suboptimal. To understand this better, note that the quality of a source code depends on the shapes of the partition cells { $\alpha^{-1}(i), i \in \mathcal{I}$ } and on varying the sizes of the cells according to the source density. When the rate is high, and either the source is uniformly distributed or the rate is measured by entropy [$H(\alpha(\mathbf{x}))$], the sizes of the cells should essentially not vary. Then, the quality depends on having cell shapes that minimize the average distance to the center of the cell.

For a given volume, a body in Euclidean space that minimizes the average distance to the center is a sphere. But spheres do not work as partition cell shapes because they do not pack together without leaving interstices. Only for a few dimensions N is the best cell shape known [2]. One such dimension is N = 2, where the hexagonal packing shown below (left) is best.

The best packings (including the hexagonal case) cannot be achieved with transform codes. Transform codes can only produce partitions into parallelepipeds, as shown for N = 2 in Fig. 3. The best parallelepipeds are cubes. We get a hint of this by comparing the two rectangular partitions of a unit-area square shown below. Both partitions have 36 cells, so every cell has the same area. The partition with square cells gives distortion $1/432 \approx 2.31 \times 10^{-3}$, while the other gives 97/31, $104 \approx 3.12 \times 10^{-3}$.

$\cdot \times \cdot \times \cdot \times \cdot$											
$[\overline{\mathcal{A}}, \overline{\mathcal{A}}, \overline{\mathcal{A}, \overline{\mathcal{A}}, \overline{\mathcal{A}}, \overline{\mathcal{A}}, \overline{\mathcal{A}}, \overline{\mathcal{A}}, \overline{\mathcal{A}}, \overline$	_		_	 	_	_	_				
$\mathbb{K} \mathbb{K} \mathbb$											
$H \cdot H \cdot H$	┤┌┤										
$H \cdot H \cdot H$	_		_								
$H \cdot H \cdot H \cdot H$											

This simple example can also be interpreted as a problem of allocating bits between the horizontal and vertical components. The "lazy" bit allocation arising from equal quantization step sizes for each component is optimal. This holds generally for high-rate entropyconstrained quantization of components with the same normalized density.

Returning now to transform choice, to get rectangular partition cells the basis vectors must be orthogonal. For square cells, when quantization step sizes are equal for each transform coefficient, the basis vectors should in addition to being orthogonal have equal lengths. When orthogonal basis vectors have unit length, the resulting transform is called an orthogonal transform. (It is regrettable that of a matrix or transform, "orthogonal" means orthonormal.)

3.1.2. Karhunen–Loève Transforms. A *Karhunen–Loève transform* (KLT) is a particular type of orthogonal

transform that depends on the covariance of the source. An orthogonal matrix T represents a KLT of \mathbf{x} if $TR_{\mathbf{x}}T^{T}$ is a diagonal matrix. The diagonal matrix $TR_{\mathbf{x}}T^{T}$ is the covariance of $\mathbf{y} = T\mathbf{x}$; thus, a KLT gives uncorrelated transform coefficients. KLT is the most commonly used name for these transforms in signal processing, communication, and information theory, recognizing the works by Karhunen and Loève [19,21]; among the other names are Hotelling transforms [16] and principal component transforms.

A KLT exists for any source because covariance matrices are symmetric, and symmetric matrices are orthogonally diagonalizable; the diagonal elements of $TR_{\mathbf{x}}T^{T}$ are the eigenvalues of $R_{\mathbf{x}}$. KLTs are not unique; any row of T can be multiplied by ± 1 without changing $TR_{\mathbf{x}}T^{T}$, and permuting the rows leaves $TR_{\mathbf{x}}T^{T}$ diagonal. If the eigenvalues of $R_{\mathbf{x}}$ are not distinct, there is additional freedom in choosing a KLT.

For an example of a KLT, consider the first-order autoregressive signal model that is popular in many branches of signal processing and communication. Under such a model, a sequence is generated as

$$\mathbf{x}[k] = \rho \mathbf{x}[k-1] + \mathbf{z}[k]$$

where k is a time index, $\mathbf{z}[k]$ is a white sequence, and $\rho \in [0, 1)$ is called the *correlation coefficient*. It is a crude but useful model for the samples along any line of a grayscale image, with $\rho \approx 0.9$.

An \mathbb{R}^N -valued source **x** can be derived from a scalar autoregressive source by forming blocks of N consecutive samples. With normalized power, the covariance matrix is given elementwise by $(R_{\mathbf{x}})_{ij} = \rho^{|i-j|}$. For any particular N and ρ , a numerical eigendecomposition method can be applied to $R_{\mathbf{x}}$ to obtain a KLT. (An analytic solution also happens to be possible [15].) For N = 8 and $\rho = 0.9$, the KLT can be depicted as follows:



Each subplot (value of k) gives a row of the transform or, equivalently, a vector in the analysis basis. This basis is superficially sinusoidal and is approximated by a discrete cosine transform (DCT) basis. There are various asymptotic equivalences between KLTs and DCTs for large N and $\rho \rightarrow 1$ [18,28]. These results are often cited in justifying the use of DCTs.

For Gaussian sources, KLTs align the partitioning with the axes of the source PDF, as shown in Fig. 3c. It appears that the forward and inverse transforms are rotations, although actually the symmetry of the source density obscures possible reflections.

3.2. The Easiest Transform Optimization

Consider a jointly Gaussian source, and assume U and T are orthogonal and $U = T^{-1}$. The Gaussian assumption is important because any linear combination of jointly Gaussian random variables is Gaussian. Thus, any analysis transform gives Gaussian transform coefficients. Then, since the transform coefficients have the same normalized density, for any reasonable set of quantizers, Eq. (2) holds with a single function g(R) describing all the transform coefficients. Orthogonality is important because orthogonal transforms preserve Euclidean lengths, which gives $d(x, \hat{x}) = d(y, \hat{y})$.

With these assumptions, for any rate and bit allocation a KLT is an optimal transform:

Theorem 1 [13]. Consider a transform coder with orthogonal analysis transform T and synthesis transform $U = T^{-1} = T^T$. Suppose that there is a single function g to describe the quantization of each transform coefficient through

$$E[(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2] = \sigma_i^2 g(R_i), \quad i = 1, 2, ..., N$$

where σ_i^2 is the variance of \mathbf{y}_i and R_i is the rate allocated to y_i . Then for any bit allocation (R_1, R_2, \ldots, R_N) there is a KLT that minimizes the distortion. In the typical case where g is nonincreasing, a KLT that gives $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$ sorted in the same order as the bit allocation minimizes the distortion.

Since it holds for any bit allocation and many families of quantizers, Theorem 1 is stronger than several earlier transform optimization results. In particular, it subsumes the low-rate results of Lyons [23] and the high-rate results that are reviewed presently.

Recall that with a high average rate of R bits per component and quantizer performance described by Eq. (8), the average distortion with optimal bit allocation is given by Eq. (9). With Gaussian transform coefficients that are optimally quantized, the distortion simplifies to

$$D = c \left(\prod_{i=1}^{N} \sigma_i^2\right)^{1/N} 2^{-2R}$$
(10)

where $c = \pi e/6$ for entropy-constrained quantization or $c = 3^{1/2}\pi/2$ for fixed-rate quantization. The choice of an orthogonal transform is thus guided by minimizing the geometric mean of the transform coefficient variances.

Theorem 2. The distortion given by Eq. (10) is minimized over all orthogonal transforms by any KLT.

Proof: Applying Hadamard's Inequality to R_y gives

$$(\det T)(\det R_{\mathbf{x}})(\det T^T) = \det R_{\mathbf{y}} \le \prod_{i=1}^N \sigma_i^2$$

Since det T = 1, the left-hand side of this inequality is invariant to the choice of *T*. Equality is achieved when a KLT is used. Thus KLTs minimize the distortion.

Equation (10) can be used to define a figure of merit called the *coding gain*. The *coding gain* of a transform is a function of its variance vector, $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$, and the variance vector without a transform, diag(R_x):

Coding gain =
$$\frac{\left(\prod_{i=1}^{N} (R_x)_{ii}\right)^{1/N}}{\left(\prod_{i=1}^{N} \sigma_i^2\right)^{1/N}}.$$

The coding gain is the factor by which the distortion is reduced because of the transform, assuming high rate and optimal bit allocation. The foregoing discussion shows that KLTs maximize coding gain. Related measures are the variance distribution, maximum reducible bits, and energy packing efficiency or energy compaction. All of these are optimized by KLTs [28].

3.3. More General Results

The results of the previous section are straightforward and Theorem 2 is well known. However, KLTs are not always optimal. With some sources, there are nonorthogonal transforms that perform better than any orthogonal transform. And, depending on the quantization, $U = T^{-1}$ is not always optimal—even for Gaussian sources. This section provides results that apply without the presumption of Gaussianity or orthogonality.

3.3.1. The Synthesis Transform U. Instead of assuming the decoder structure shown in the bottom of Fig. 1, let us consider for a moment the best way to decode given only the encoding structure of a transform code. The analysis transform followed by quantization induces some partition of \mathbb{R}^N , and the best decoding is to associate with each partition cell its centroid. Generally, this decoder cannot be realized with a linear transform applied to \hat{y} . For one thing, some scalar quantizer decoder β_i could be designed in a plainly wrong way; then it would take an extraordinary (nonlinear) effort to fix the estimates.

The difficulty is actually more dramatic because even if the β_i mappings are optimal, the synthesis transform T^{-1} applied to \hat{y} will seldom give optimal estimates. In fact, unless the transform coefficients are independent, there may be a *linear transform* better suited to the reconstruction than T^{-1} .

Theorem 3 [12]. In a transform coder with invertible analysis transform T, suppose that the transform coefficients are independent. If the component quantizers reconstruct to centroids, then $U = T^{-1}$ gives centroid reconstructions for the partition induced by the encoder. As a further consequence, T^{-1} is the optimal synthesis transform.

Examples where the lack of independence of transform coefficients or the absence of optimal scalar decoding makes T^{-1} a suboptimal synthesis transform are given in Ref. 12.

3.3.2. The Analysis Transform T. Now consider the optimization of T under the assumption that $U = T^{-1}$. The first result is a counterpart to Theorem 2. Instead of requiring orthogonal transforms and finding uncorrelated transform coefficients to be best, it requires independent transform coefficients and finds orthogonal basis vectors to be best. It does not require a Gaussian source; however, it is only for Gaussian sources that R_y being diagonal implies that the transform coefficients are independent.

Theorem 4 [12]. Consider a transform coder in which analysis transform T produces independent transform coefficients, the synthesis transform is T^{-1} , and the component quantizers reconstruct to their respective centroids. To minimize the MSE distortion, it is sufficient to consider transforms with orthogonal rows, that is, Tsuch that TT^{T} is a diagonal matrix.

The scaling of a row of T is generally irrelevant because it can be completely absorbed in the quantizer for the corresponding transform coefficient. Thus, Theorem 4 implies furthermore that it suffices to consider orthogonal transforms that produce independent transform coefficients. Together with Theorem 1, it still falls short of showing that a KLT is necessarily an optimal transform — even for a Gaussian source.

Heuristically, independence of transform coefficients seems desirable because otherwise dependencies that would make it easier to code the source are "wasted." Orthogonality is beneficial for having good partition cell shapes. A firm result along these lines requires high resolution analysis:

Theorem 5 [12]. Consider a high-rate transform coding system employing entropy-constrained uniform quantization. A transform with orthogonal rows that produces independent transform coefficients is optimal when such a transform exists. Furthermore, the norm of the *i*th row divided by the *i*th quantizer step size is optimally a constant. Thus, normalizing the rows to have an orthogonal transform and using equal quantizer step sizes is optimal.

For Gaussian sources there is always an orthogonal transform that produces independent transform coefficients—the KLT. For some other sources there are only nonorthogonal transforms that give independent transform coefficients, but for most sources there is no linear transform that does so. Is it more important to have orthogonality or independent transform coefficients? Examples in Ref. 12 demonstrate that there is no unequivocal answer. However, in practical situations the point is moot because it is not possible to assure independence of transform coefficients. Thus orthogonal transforms are almost always used.

4. DEPARTURES FROM THE STANDARD MODEL

4.1. Scalar and Vector Entropy Coding

Practical transform coders differ from the standard model in many ways. One particular change has significant implications for the relevance of the conventional analysis and has lead to new theoretical developments: Transform coefficients are often not entropy coded independently.

Allowing transform coefficients to be entropy coded together, as it is drawn in Fig. 1, throws the theory into disarray. Most significantly, it eliminates the incentive to have independent transform coefficients. As for the particulars of the theory, it also destroys the concept of bit allocation because bits are shared among transform coefficients.

The status of the conventional theory is not quite so dire, however, because the complexity of unconstrained joint entropy coding of the transform coefficients is prohibitive. Assuming an alphabet size of K for each scalar component, an entropy code for vectors of length N has K^N codewords. The problems of storing this codebook and searching for desired entries prevent large values of N from being feasible. Entropy coding without explicit storage of the codewords—as, for example, in arithmetic coding—is also difficult because of the number of symbol probabilities that must be known or estimated.

Analogous to constrained lossy source codes, joint entropy codes for transform coefficients are usually constrained in some way. In the original JPEG standard [27], the joint coding is limited to transform coefficients with quantized values equal to zero. This type of joint coding does not eliminate the optimality of the KLT (for a Gaussian source); in fact, it makes it even more important for a transform to give a large fraction of coefficients with small magnitude. The empirical fact that wavelet transforms have this property for natural images (more abstractly, for piecewise smooth functions) is a key to their current popularity.

Returning to the choice of a transform assuming joint entropy coding, the high-rate case gives an interesting result—quantizing in any orthogonal analysis basis and using uniform quantizers with equal step sizes is optimal. All the meaningful work is done by the entropy coder. Given that the transform has no effect on the performance, it can be eliminated. There is still room for improvement, however.

Producing transform coefficients that are independent allows for the use of scalar entropy codes, with the attendant reduction in complexity, without any loss in performance. A transform applied to the quantizer outputs, as shown in Fig. 4, can be used to achieve or approximate this. For Gaussian sources, it is even possible to design the transform so that the quantized transform coefficients have approximately the same distribution, in addition to being approximately independent. Then the same scalar entropy code can be applied to each transform



Figure 4. An alternative transform coding structure introduced in Ref. 10. The transform *T* operates on a vector of discrete quantizer indices instead of on the continuous-valued source vector, as in the standard model. Scalar entropy coding is explicitly indicated. For a Gaussian source, a further simplification with $\gamma_1 = \gamma_2 = \cdots = \gamma_N$ can be made with no loss in performance.

coefficient [10]. Some of the lossless codes used in practice include transforms, but they are not optimized for a single scalar entropy code.

4.2. Transmission with Losses

When source coding is separated completely from the underlying communication problem, it is assumed, as we have done so far, that bits are communicated without loss or error from the encoder to the decoder. Transforms may also be used when some data is lost in transmission though the design objectives for the transform may be changed. Some techniques for these situations have been described elsewhere [11].

5. HISTORICAL NOTES

Transform coding was invented as a method for conserving bandwidth in the transmission of signals output by the analysis unit of a 10-channel vocoder ("voice coder") [4]. These correlated, continuous-time, continuous-amplitude signals represented estimates, local in time, of the power in ten contiguous frequency bands. By adding modulated versions of these power signals, the synthesis unit resynthesized speech. The vocoder was publicized through the demonstration of a related device, called the *Voder*, at the 1939 World's Fair.

Kramer and Mathews [20] showed that the total bandwidth necessary to transmit the signals with a prescribed fidelity can be reduced by transmitting an appropriate set of linear combinations of the signals instead of the signals themselves. Assuming Gaussian signals, KLTs are optimal for this application.

The technique of Kramer and Mathews is not source coding because it does not involve discretization. Thus, one could ascribe a later birth to transform coding. Huang and Schultheiss [17] introduced the structure shown in the bottom of Fig. 1, which we have referred to as the *standard model*. They studied the coding of Gaussian sources while assuming independent transform coefficients and optimal fixed-rate scalar quantization. First they showed that $U = T^{-1}$ is optimal and then that T should have orthogonal rows. These results are subsumed by Theorems 3 and 4. They also obtained high-rate bit allocation results.

6. SUMMARY

The theory of source coding tells us that performance is best when large blocks of data are used. But this same theory suggests codes that are too difficult to use—because of storage, running time, or both—if the block length is large. Transform codes alleviate this dilemma. They perform well, although not optimally, but are simple enough to apply with very large block lengths.

Transform codes are easy to implement because of a divide-and-conquer strategy; the transform exploits dependencies in the data so that the quantization and entropy coding can be simple. As for the choice of the transform, two qualitative intuitions arise from high-resolution transform coding theory: (1) try to have independent transform coefficients; and (2) use orthogonal transforms. Sometimes you can only have one or the other, but for jointly Gaussian sources this works perfectly because you can have both: the transform can produce independent transform coefficients, and then little is lost by using scalar quantization and scalar entropy coding.

As with source coding generally, it is hard to make use of the theory of transform coding with real-world signals. Nevertheless, the principles of transform coding certainly do apply, as evidenced by the dominance of transform codes in audio, image, and video compression.

Acknowledgment

Artical adapted, with permission, from "Theoretical Foundations of Transform Coding," *IEEE Signal Processing Magazine*, Vol. 18, No. 5, pp. 9–21, September 2001. © 2001 IEEE. Condensed from [12].

BIOGRAPHY

Vivek K. Goyal received his B.S. degree in mathematics and his B.S.E. in electrical engineering (both with highest distinction), in 1993, from the University of Iowa, Iowa City. He received his M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1995 and 1998, respectively. In 1998 he received the Eliahu Jury Award of the University of California, Berkeley, awarded to a graduate student or recent alumnus for outstanding achievement in systems, communications, control, or signal processing.

Dr. Goyal was a research assistant in the Laboratoire de Communications Audiovisuelles at Ecole Polytechnique Federale de Lausanne, Switzerland, in 1996. He worked in the Mathematics of Communications Research Department of Lucent Technologies Bell Laboratories as an intern in 1997 and again as a member of technical staff from 1998 to 2001. He is currently a senior research engineer for Digital Fountain, Inc., Fremont, California. Dr. Goyal is a member of Phi Beta Kappa, Tau Beta Pi, Sigma Xi, Eta Kappa Nu, IEEE, and SIAM. He serves on the program committee of the IEEE Data Compression Conference. His research interests include source coding theory, quantization theory, and practical, robust network content delivery.

BIBLIOGRAPHY

- T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- J. H. Conway and N. J. A. Sloane, Sphere Packings, Lattices and Groups, Vol. 290 of Grundlehren der mathematischen Wissenschaften, 3rd ed., Springer-Verlag, New York, 1998.
- T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
- H. W. Dudley, The vocoder, *Bell Lab. Rec.* 18: 122–126 (Dec. 1939).
- M. Effros, Optimal modeling for complex system design, *IEEE Signal Process. Mag.* 15(6): 51–73 (Nov. 1998).
- N. Farvardin and J. W. Modestino, Optimum quantizer performance for a class of non-Gaussian memoryless sources, *IEEE Trans. Inform. Theory* **IT-30**(3): 485–497 (May 1984).
- A. Gersho, Asymptotically optimal block quantization, *IEEE Trans. Inform. Theory* IT-25(4): 373–380 (July 1979).
- A. Gersho and R. M. Gray, Vector Quantization and Signal Compression, Kluwer, Boston, 1992.
- H. Gish and J. P. Pierce, Asymptotically efficient quantizing, IEEE Trans. Inform. Theory IT-14(5): 676-683 (Sep. 1968).
- V. K. Goyal, Transform coding with integer-to-integer transforms, *IEEE Trans. Inform. Theory* 46(2): 465–473 (March 2000).
- V. K. Goyal, Multiple description coding: Compression meets the network, *IEEE Signal Process. Mag.* 18(5): 74–93 (Sept. 2001).
- 12. V. K. Goyal, Single and Multiple Description Transform Coding with Bases and Frames, SIAM, 2002.
- V. K. Goyal, J. Zhuang, and M. Vetterli, Transform coding with backward adaptive updates, *IEEE Trans. Inform. Theory* 46(4): 1623–1633 (July 2000).
- R. M. Gray and D. L. Neuhoff, Quantization, *IEEE Trans. Inform. Theory* 44(6): 2325–2383 (Oct. 1998).
- 15. U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, Univ. California Press, Berkeley, CA, 1958.
- H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24: 417–441, 498–520 (1933).

- J. J. Y. Huang and P. M. Schultheiss, Block quantization of correlated Gaussian random variables, *IEEE Trans. Commun. Syst.* 11: 289–296 (Sept. 1963).
- A. K. Jain, A sinusoidal family of unitary transforms, *IEEE Trans. Pattern Anal. Mach. Int.* PAMI-1(4): 356–365 (Oct. 1979).
- K. Karhunen, Über lineare methoden in der Wahrscheinlichkeitsrechnung, Ann. Acad. Sci. Fenn., Ser. A.1.: Math.-Phys. 37: 3-79 (1947).
- H. P. Kramer and M. V. Mathews, A linear coding for transmitting a set of correlated signals, *IRE Trans. Inform. Theory* 23(3): 41-46 (Sept. 1956).
- M. Loève, Fonctions aleatoires de seconde ordre, in P. Levy, ed., Processus Stochastiques et Mouvement Brownien, Gauthier-Villars, Paris, 1948.
- D. G. Luenberger, Optimization by Vector Space Methods, Wiley, New York, 1969.
- D. F. Lyons III, Fundamental Limits of Low-Rate Transform Codes, Ph.D. thesis, Univ. Michigan, 1992.
- N. Moayeri and D. L. Neuhoff, Time-memory tradeoffs in vector quantizer codebook searching based on decision trees, *IEEE Trans. Speech Audio Process.* 2(4): 490–506 (Oct. 1994).
- A. Ortega and K. Ramchandran, Rate-distortion methods for image and video compression, *IEEE Signal Process. Mag.* 15(6): 23-50 (Nov. 1998).
- A. Papoulis, Probability, Random Variables, and Stochastic Processes, 3rd ed., McGraw-Hill, New York, 1991.
- W. B. Pennebaker and J. L. Mitchell, JPEG Still Image Data Compression Standard, Van Nostrand Reinhold, New York, 1993.
- K. R. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, Applications, Academic Press, San Diego, CA, 1990.
- A. Segall, Bit allocation and encoding for vector sources, *IEEE Trans. Inform. Theory* IT-22(2): 162–169 (March 1976).
- Y. Shoham and A. Gersho, Efficient bit allocation for an arbitrary set of quantizers, *IEEE Trans. Acoust. Speech* Signal Process. 36(9): 1445-1453 (Sept. 1988).

TRANSMISSION CONTROL PROTOCOL

JAMES AWEYA Nortel Networks Ottawa, Ontario, Canada

1. INTRODUCTION

The Internet, which has become a global communication network, uses packet switching techniques to enable its attached devices — personal computers, workstations, servers, and wireless devices — to exchange information. The information is encoded as long strings of bits called *packets*. In order to achieve the transfer of packets between the attached devices, certain rules about packet format and processing must be followed. These rules are called *protocols*, and the suite of protocols used by the Internet is the TCP/IP. The Internet's marked success has made the TCP/IP protocol suite the most ubiquitous tool for computer networking. Hence, the most widely used transport protocols today are Transmission Control Protocol (TCP) [1] and its companion transport protocol, the User Datagram Protocol (UDP) [2]. Although a number of other transport protocols have been developed or proposed [3], the success of the Internet has made TCP and UDP the dominant transport protocols for networking.

The Internet Protocol (IP) is fundamental to Internet addressing and routing, while the transport protocols provide an end-to-end connection between application processes running in the source and destination end devices. The application processes are identified by *port numbers*. Both TCP and UDP run on top of IP and build on the connectionless datagram services provided by IP. TCP provides a reliable, connection-oriented, bytestream service between applications. UDP, on the other hand, provides service without reliability guarantees. It is a lightweight protocol that allows applications to make direct use of the unreliable datagram service provided by the underlying IP service. UDP is basically an interface to IP, adding little more than multiplexing/demultiplexing (port numbers) and optional data integrity service. By not providing reliability service, UDP's overhead is significantly less than that of TCP. Although UDP includes an optional checksum, incoming datagrams with checksum errors are silently discarded and no error message is generated, while valid datagrams are passed to the application.

The reliable transport service provided by TCP is used by most Internet applications, including interactive Telnet, file transfer, electronic mail, and Webpage access via the Hypertext Transfer Protocol (HTTP). In this article, we look at the basic design and operation of TCP, particularly the main features that make TCP a reliable transport protocol.

2. BASIC TCP FEATURES

TCP provides reliability and ensures end-to-end delivery of data by adding services on top of IP. IP is connectionless and does not guarantee delivery of packets. TCP provides a full-duplex (i.e., can carry data in both directions), virtual circuit connection between two applications communicating with each other. The applications communicate across the TCP connection by exchanging a stream of 8-bit bytes in each direction. TCP groups a set of bytes that need to be sent into a message segment that is passed to IP. Message segments can be of arbitrary length, but for reasons of efficiency in managing messages, message segments can be limited by a maximum segment size (MSS) that each end has the option of announcing when a connection is established. Applications that use TCP send data in whatever size is convenient for sending. Applications can send data to TCP a few bytes (as little as one byte) or several kilobytes at a time. TCP buffers these data and sends these bytes either as single message segment or as several smaller message segments. Ultimately, the messages are sent in IP datagrams that are limited by the maximum transmission unit (MTU) of a network interface.

TCP treats the actual data it sends as an unstructured stream of bytes. It does not contain any facility to

superimpose an application-dependent structure on the data. For example, an application cannot instruct TCP to treat the data as a set of records in a database and to send one record at a time. Any such structuring must be handled by the application that communicates using TCP. Because TCP sends data as a stream of bytes, there is no real end-of-message marker in the datastream.

TCP keeps track of each byte that is sent/received. It has no inherent notion of a block of data, unlike other transport protocols, which typically keep track of the Transport Protocol Data Unit (TPDU) number and not the byte number. TCP numbers each byte that it sends and the number assigned to each byte is called the *sequence number*. This number is necessary to ensure that the bytes are delivered to the application at the receiving end in the order in which they are sent. This process is called *sequencing* of the bytes.

In order to provide a reliable service, TCP must recover from data that is lost, damaged, duplicated, or delivered out of order by IP. TCP achieves this using the positive acknowledgment retransmission (PAR) scheme. TCP implements PAR by assigning a sequence number to each byte that is transmitted and requiring a positive acknowledgment (ACK) from the receiving TCP module. If the ACK is not received within a timeout interval, the data are retransmitted. At the receiver TCP module, the sequence numbers are used to correctly order segments that may have arrived out of order and to eliminate duplicates. Corruption of data is detected by using a checksum field in the TCP header. Data segments that are received with a bad checksum field are discarded.

Since Internet devices can send and receive TCP data segments at different rates because of differences in processor and network bandwidth, it is quite possible for a sending device to send data at a much faster rate than the receiver can handle. TCP implements a flow control mechanism that controls the amount of data sent by the sender. TCP uses a *sliding window* mechanism for implementing flow control. The goal of the sliding window mechanism is to keep the channel full of data and to reduce to a minimum the delays experienced in waiting for acknowledgments.

TCP enables many application processes within a single device to use the TCP services simultaneously; this is termed TCP *multiplexing*. Those processes that may be communicating over the same network interface are identified by the IP address of the network interface. TCP associates a port number value for applications that use TCP. This association enables several connections to exist between application processes on devices because each connection uses a different pair of port numbers. The binding of ports to application processes is handled independently by a device.

TCP applications must establish a connection between them before they can exchange data. A TCP connection identifies the endpoints involved in the connection. An *endpoint* is defined as a pair that includes the unique IP address of the network interface over which the application communicates, and the port number that identifies the application. The TCP connection is identified by the parameters of both endpoints as follows: {IP address1, port number1, IP address2, port number2}. A connection is fully specified by the pair of endpoints. These parameters make it possible to have several application processes connected to the same endpoint. A local endpoint can participate in many connections to different foreign endpoints.

3. TCP MESSAGE FORMAT

A TCP message segment consists of a header part and an optional data part. The header part, which can be up to 60 bytes long, further comprises a fixed section of 20 bytes to carry 15 fields, and an optional section that can carry up to 40 bytes of TCP options.

3.1. Fixed Header Fields

The TCP header format is shown in Fig. 1. The header has a normal size of 20 bytes, unless TCP options are present.

The pair of 16-bit source and destination *port number* fields is used to identify the endpoint applications of the TCP connection. The source and destination IP addresses, and source and destination port numbers uniquely identify a TCP connection. Some port numbers are well-known port numbers; others have been registered, and still others are dynamically assigned.

The 32-bit sequence number identifies the first byte of the data in a message segment and is sent in the TCP header for that segment. If the SYN flag field is set to 1, the sequence number field defines the *initial sequence number* (ISN) to be used for that session, and the first data offset is ISN+1. The ISN does not take a value of 1 for a new TCP connection. The value of the ISN selected is intended to prevent delayed data from an old connection (i.e., old sequence numbers that already may have been assigned to data that are in transit on the network) from being incorrectly interpreted as being valid within the current connection.

Since TCP transmissions are full-duplex, a TCP module is both a sender and receiver of data. When a message is sent by the receiver to the sender, it also carries a 32-bit *acknowledgment number*, which indicates the sequence number of the next byte expected by the receiver. That is, the acknowledgment number is 1 plus the sequence number of the last successfully received byte of data. TCP acknowledgments are cumulative; that is, a single acknowledgment can be used to acknowledge a number of prior TCP message segments. Transmission by TCP is made reliable via the use of sequence numbers and acknowledgment numbers.

The 4-bit *data offset* (or *header length*) field is the number of 32-bit words in the TCP header. This field is needed because the TCP options field could be variable in length. Without TCP options, the data offset field is 20 bytes (5 words). The 4-bit field limits the TCP header to 60 bytes. The 6-bit field marked *reserved* is reserved for future use.

The six 1-bit flag fields in the TCP header serve the following purposes:

- Urgent Pointer Flag (URG) When the URG flag is set, it indicates that the *urgent pointer* is valid. Using these two fields, TCP allows one end of the connection to inform the other end that urgent data have been placed in the normal data stream. This feature requires that when urgent data are found, the receiving TCP should notify whatever application is associated with the connection to go into "urgent mode." After all urgent data have been consumed, the application returns to normal operation.
- Acknowledgment Flag (ACK) When the ACK flag is set, it indicates that the acknowledgment number field is valid.
- Push Flag (PSH) When the PSH flag is set, it tells TCP immediately to deliver data for this message to the upper layer process.
- Reset Flag (RST) The RST flag is used to reset the connection. When an RST is received in a TCP segment, the receiver must respond by immediately terminating the connection. A reset causes the immediate release of the connection and its resources. Sending a RST is not the normal way to close a TCP connection. The normal way where a FIN is sent after all previously queued data has been successfully sent



Figure 1. TCP header format.

is sometimes called an *orderly release*. Aborting a connection by sending a RST instead of a FIN is sometimes referred to as an *abortive release*. Other than for an abortive release, one common reason for generating a reset is when a connection request arrives at a destination port that is not in use by an application process.

- Synchronize Sequence Numbers Flag (SYN) The SYN flag is used to indicate the opening of a TCP connection.
- Finish Flag (FIN) The FIN flag is used to terminate the connection.

The 16-bit *window size* field is used to implement flow control and reflects the amount of buffer space available for new data at the receiver. The receiving TCP reports a window to the sending TCP, and this window specifies the number of bytes, starting with the acknowledgment number, that the receiving TCP is currently prepared to receive. The 16-bit window limits the window to 65,535 bytes, but a window scale option allows this value to be scaled to provide larger windows.

The 16-bit checksum, which is mandatory, is used to verify the integrity of the TCP header as well as the data. The TCP checksum is an end-to-end checksum. It is calculated by the sender and then verified by the receiver. The checksum is the one's complement of the one'scomplement sum of all the 16-bit words in the TCP packet. A 12-byte pseudoheader (see Fig. 2) is prepended to the TCP header for checksum computation. The pseudoheader is used to identify whether the packet has arrived at the correct destination. The pseudoheader gives the TCP protection against misrouted segments. The TCP segment can be an odd number of bytes, while the checksum algorithm adds 16-bit words. So a pad byte of 0 is appended to the end, if necessary, just for checksum computation. The 16-bit TCP length field in Fig. 2 (which is a computed quantity and appears twice in the checksum computation) is the TCP header length plus the data length in bytes. This field does not count the 12 bytes of the pseudoheader.

The URG and *urgent pointer* fields constitute a mechanism by which TCP marks urgent data when transmitting segments. The 16-bit urgent pointer is a positive offset when added to the sequence number field of the segment, indicates the sequence number of the last

byte in the sequence of urgent data. This feature enables the sender to send interrupt signals to the receiver and prevents these signals from ending up in the normal data queue at the receiver.

3.2. TCP Options

Many options can be specified in the *TCP options* field up to a maximum of 40 bytes as allowed by the 4-bit data offset field (which limits the TCP header to 60 bytes). A number of TCP options have been defined [1,4,5]; however, the current options relevant to TCP performance are

- Maximum Segment Size (MSS) Option. The MSS option [1] is used only when a connection is being established, and appears only in the initial SYN segment used to open the connection. A TCP sender uses this option to inform the remote end of the largest unit of information or maximum segment size it is willing to receive on the TCP connection. The setting of the MSS option can be up to the MTU of the outgoing interface minus the size of the TCP and IP headers. The MSS option when used together with path MTU discovery [6] allows for the establishment of a segment size that can be sent across the path between two hosts without fragmentation. Fragmentation is undesirable because if one fragment is lost, TCP will timeout and retransmit the entire TCP segment.
- Window-Scale Option. This option allows TCP to use window sizes that can operate efficiently over largebandwidth-delay networks. The 16-bit window size field in the TCP header limits the window to 65,535 bytes. However, most networks, in particular high-speed networks and networks with satellite links, will require a much higher window than this for maximum TCP throughput. The window-scale option effectively increases the size of the window to a 30-bit field, but only the most significant 16 bits of the value are transmitted. This allows larger windows, up to 2³⁰ bytes, to be transmitted [4]. This option can only be sent in a SYN segment at the start of the TCP connection.
- SACK Option. The selective acknowledgment (SACK) option in TCP is defined in RFC 2018 [5]. This option is used to modify the acknowledgment



Figure 2. Fields used for TCP checksum computation.

behavior of TCP. The default behavior is to acknowledge the highest sequence number of bytes received in order. The SACK option allows for more robust operation when multiple segments are lost from a single window of data. It enables a TCP receiver to inform the sender of what specific segments were lost (selective acknowledgment) so that the TCP sender can retransmit them. Thus, when faced with multiple segment losses in a single window, TCP SACK enables a sender to continue to transmit segments (retransmissions and new segments) without entering into a time-consuming slow-start phase.

The *data* portion of the TCP segment is optional; a TCP segment does not necessarily need to carry user data. In particular, when a connection is established, and when a connection is terminated, segments with TCP header and possibly with options, are exchanged. The TCP segment used to acknowledge received data when there are no data to be transmitted in that direction consists of a header only.

4. TCP OPERATION

Because IP provides no sequencing or acknowledgment of data and is connectionless, the tasks of connection establishment and termination, reliability in data transfer, data sequencing, and flow control are given to TCP [1]. These operational features of TCP are discussed in this section.

4.1. TCP Connection Establishment

Two applications using TCP must establish a *connection* with each other before they can exchange data. A TCP connection is established using a *three-way handshake*, which ensures that both ends of the connection have a clear understanding of the initial sequence number, and possibly, the TCP options of the remote end. The three steps involved in the three-way handshake are summarized as follows:

Step 1—endpoint 1 sends a SYN segment (with flag fields SYN = 1, ACK = 0) specifying its initial

sequence number, x, and the port number to endpoint 2.

- Step 2—endpoint 2 responds with a SYN segment (with flag fields SYN = 1, ACK = 1) containing its own initial sequence number, y, and an acknowledgment (ACK) of the received initial sequence number (acknowledgment number is 1 greater then x).
- Step 3—endpoint 1 acknowledges the received remote sequence number by sending a segment (with flag fields SYN = 0, ACK = 1) containing the acknowledgment number y + 1.

The three steps are shown in Fig. 3. It takes three segments to establish a connection, and 1.5 round-trip times (RTTs) for the two end systems to synchronize state before data exchange. An *active open* is said to be performed by the endpoint that sends the first SYN, while a *passive open* is performed by the other end that receives this first SYN and sends the next SYN.

After a TCP connection has been established, the ACK flag is always set to 1 to indicate that the acknowledgment number field is valid. Data transmission then takes place with messages being exchanged by both sides in a full-duplex fashion until the TCP connection is ready to be closed. When an application process notifies TCP that it has no more data to send, TCP will close the connection in *the sending direction*. Thus, each direction of data flow must be shut down independently since a TCP connection is full-duplex, and can be viewed as containing two independent stream transfers, one going in each direction. *TCP half-close* refers to the ability of one end of a TCP connection to terminate data transfer (except the output of acknowledgment segments) while still receiving data from the other half.

The FIN control flag is issued when a TCP connection is ready to close. One end sends a segment with the FIN flag set to close its half of the connection when it has finished sending data. The receiving TCP acknowledges the FIN segment and notifies the application process at its end that no more data will be delivered to it. However, data can continue to flow in the other direction (i.e., other half of the connection) until that direction of data flow is closed. Acknowledgment segments continue to flow back





2608 TRANSMISSION CONTROL PROTOCOL



Figure 4. TCP connection termination: (a) normal connection termination; (b) TCP half-close.

to the sender even after a connection is half-closed. The end that sends the first FIN is said to perform an *active close* and the end that receives this FIN performs a *passive close*. The connection termination procedure is illustrated in Fig. 4. When both directions have been closed, then data completely stop flowing in both directions. In practice, few TCP applications use the half-close feature.

4.2. Data Transfer

TCP can handle both interactive data transfers and bulk data transfers. However, the methods of flow control and data acknowledgment differ between these two application areas. **4.2.1.** Interactive Data Transfer. *Telnet* and *rlogin* are examples of applications that generate interactive data. Interactive applications typically send data in very small units. For example, with rlogin, only one byte is sent in a segment. Obviously, this type of operation translates into considerable overhead, since this one byte generates a 41-byte packet: 20 bytes for the TCP header and 20 bytes for IP header. Some performance improvements in data exchange can be obtained through the use of *delayed acknowledgments* [7] and *ACK piggybacking*, where the receiver holds back the ACK for a brief time (*delayed acknowledgment*) and attempts to piggyback the ACK onto data going back to the sender (*ACK piggybacking*). This helps reduce the number of segments sent into the network

since some interactive applications, such as rlogin, require that the receiver echo back the character (byte) that is received. Some operations of interactive data exchange are illustrated in Fig. 5.

For short-delay links, the high-overhead packets (called *tinygrams*) do not pose significant performance problems in the interactive data exchange. However, for slow large-delay links, these high-overhead packets can be a source of congestion traffic. A mechanism commonly called the *Nagle algorithm* was proposed in RFC 896 [8] to reduce the number of these small packets. The Nagle algorithm



Figure 5. Interactive data exchange: (a) without delayed ACK; (b) with delayed ACK.

allows only one small segment to be outstanding at a time without acknowledgment. If more small segments are generated while awaiting the acknowledgment for the first one, then these segments are coalesced into a larger segment and sent when the acknowledgment arrives. On short-delay links where the return of ACKs is faster, the algorithm has negligible impact on data transfer. However, on slow links, where it is desirable to reduce the number of small packets, fewer such packets are transmitted. Although the Nagle algorithm can improve data transfer efficiency by transmitting packets with higher payload-to-header ratios, it may not be appropriate for all interactive applications, especially applications that are jitter sensitive. Jitter sensitive interactive applications typically demand that the small messages they generate be delivered without delay to provide real-time feedback to users. Using the Nagle algorithm can result in an increase in the session jitter by up to a round-trip time (RTT) interval. Interactive applications that are jitter-sensitive typically disable this algorithm.

4.2.2. Bulk Data Transfer. This section describes the mechanisms used by TCP for bulk data transfer. These mechanisms allow a sender to transmit multiple data segments before receiving an acknowledgment for those segments. They also allow a TCP sender to respond to congestion and data loss at any point along the network path between the sender and the receiver.

4.2.2.1. Sliding-Window Protocol. TCP uses a slidingwindow mechanism for bulk data transfer (see Fig. 6). The stream of data has a sequence number assigned at the byte level. The receiver TCP module sends back to the sender an acknowledgment that indicates the range of acceptable sequence numbers beyond the last segment successfully received. This range of acceptable sequence numbers is called a *window*. The window, therefore, indicates the number of bytes that the sender can transmit before receiving further permission.

In Fig. 6, bytes that are to the left of the window range have already been sent and acknowledged. Bytes in the window range can be sent without any delay. Some of the bytes in the window range may already have been sent, but they have not been acknowledged. Other bytes may be waiting to be sent. Bytes that are to the right of the window range have not been sent. These bytes can be sent only when they fall in the window range.

The left edge of the window is the lowest numbered byte that has not been acknowledged. The window can advance; that is, the left edge of the window can move to the right when an acknowledgment is received for the data that have been sent. The window size (called the



Figure 6. TCP sliding-window flow control.

"receiver's advertised window size") reflects the amount of buffer space available for new data at the receiver. When the receiving TCP process frees up buffer space by reading acknowledged data, the right edge of the window moves to the right. If this buffer space size shrinks because the receiver is being overrun, the receiver will send back a smaller window size. In the extreme case, it is possible for the receiver to send a window size of only one byte (instead of waiting until a larger window could be sent), which means that only one byte can be sent. This situation is referred to as the "silly-window syndrome," and most TCP implementations take special measures to avoid it. The silly-window syndrome can also be caused by the sender when it generates and transmits small amounts of data at a time instead of waiting to generate enough data that can be sent in larger segments.

When a TCP module sends back a window size of zero, it indicates to the sender that its buffers are full and no additional data should be sent. This happens when the left edge of the window reaches the right edge. The sliding window mechanism allows TCP to shrink the window size when the receiver experiences congestion of data and to expand the window size as the congestion problem clears.

It takes about one round-trip time (RTT) between sender and receiver for a sender to receive acknowledgment to data sent. For maximum transfer efficiency, the TCP sender must be able to completely fill the data pipe of the connection with data. Thus, the size of the window offered by the TCP receiver must be large, since it limits how much the sender can transmit. Making the advertised window no smaller than the *bandwidth-delay product* (i.e., capacity in bytes or bits) of the connection path allows maximum data transfer. This results in a window size of

Window size \geq bandwidth (bps) \times round-trip time (sc)

4.2.2.2. TCP Congestion Control Protocols. We have seen above that either the bandwidth or the RTT of a TCP connection can affect its window size. We have also seen that the larger the window size, the more data can be sent across the connection. However, given a window size, a TCP sender adopts a more controlled behavior in utilizing that window size. This is because a TCP sender initially has no idea of the available network capacity. Also, if a TCP sender commences by injecting a full window size into the network, particularly using a large window size, then there is a strong chance that much of this burst of data would be lost because of transient congestion in the network. Congestion can occur in the network nodes (e.g., routers) when data arrive on a large-bandwidth link and exit on a smaller-bandwidth link. Congestion can also occur when multiple input datastreams converge at a network node whose output bandwidth is less than the sum of the inputs. For these reasons, TCP starts by transmitting small amounts of data noting that these have a better chance of getting through to the receiver, and then probing the network with increasing amounts of data until the network shows signs of congestion. When TCP determines that the network is showing signs of congestion, it reduces its sending rate and then resumes the probing for additional bandwidth.

This seemingly seesaw behavior of TCP is its way of locating the point of equilibrium of maximum network efficiency where its sending rate is maximized just prior to the onset of sustained data loss. The bandwidth probing action of the TCP sender also helps it locate the point at which its data transmission rate is synchronized to the data extraction rate of the receiver. At this point, the rate of return of ACKs to the sender is identical to the rate of transmission of data segments. This is called the *selfclocking* behavior of TCP. This is so called because the receiver can only generate ACKs when data arrive, and the rate of arrival of the ACKs at the sender identifies the arrival rate of the segments at the receiver.

Current TCP implementations contain a number of mechanisms aimed at controlling data transmission and network congestion [9-11]. In addition to the receiver's advertised window (i.e., the buffer size advertised in acknowledgments), a TCP sender maintains a second window called the congestion window (cwnd). The cwnd tracks the maximum amount of data a sender can transmit before an ACK is required. TCP data transfer commences in a phase called *slow start*. The limit of the slow-start process is maintained in the state variable *ssthresh*. The initial *cwnd* used by the sender at the start of this phase is set to a segment size that is equal to any one of the following: the MSS obtained during the three-way handshake, the segment size resulting from the use of the path MTU discovery protocol, the MTU of the sending TCP interface less the TCP and IP headers, or the default segment size of 536 bytes. TCP then sends a segment not exceeding this first window size and waits for the corresponding acknowledgment.

When the acknowledgment is received, *cwnd* is incremented from 1 to 2, which means two segments can be sent. When each of these two segments is acknowledged, the congestion window *cwnd* is increased to four. Each time an ACK is received, *cwnd* is increased by one segment. This essentially leads to an exponential increase in the congestion window in slow start if the receiver sends an ACK for every segment received. The rate of congestion window increase would be slightly lower if the receiver implements delayed ACKs; nevertheless, the rate of increase is rapid. TCP maintains the congestion window *cwnd* in bytes, but always increments it by the segment size in slow start. Each time, the actual window size that is used by the TCP sender is the smaller of the congestion window and the receiver's advertised window.

The slow-start phase is terminated when *cwnd* equals the value of *ssthresh*, the receiver's advertised window, or when congestion (or packet loss) occurs. TCP then goes into what is termed the *congestion avoidance* phase, where it takes a more conservative approach in probing for network bandwidth. In this phase, the congestion window *cwnd* is incremented by 1/*cwnd* each time an ACK is received. Congestion avoidance thus increases the TCP sending rate in a linear fashion. In other words, the *cwnd* is incremented by at most one segment per round-trip time (regardless of how many ACKs are received) until congestion is detected, or the receiver's advertised window is reached. Unlike slow start's exponential increase, where *cwnd* is incremented by the number of ACKs received in a



Figure 7. The slow-start and congestion avoidance phases.

RTT, congestion avoidance adopts an additive increase process. The slow-start and the congestion avoidance phases are illustrated in Fig. 7. The *ssthresh* is initially set to the receiver's maximum window size. However, when congestion occurs, the *ssthresh* is set to one-half of the current window size (i.e., the minimum of *cwnd* and the receiver's advertised window, but at least two segments). This provides TCP with the best guess of the point where network congestion could occur in the future.

TCP uses the reception of *duplicate* ACKs or the expiration of a time (*timeout*) to detect when a data segment loss occurs. The behavior of TCP when a segment is lost or received out of order can be explained as follows:

- 1. When a single segment is lost in a sequence of segments, the successful reception of segments following the lost segment will cause the receiver to issue a duplicate ACK for each successfully received segment.
- 2. When the TCP receiver gets a segment with an out-of-order sequence number value, the receiver is required to generate an immediate ACK of the highest in-order data byte received (a duplicate ACK of an earlier transmission). The purpose of the duplicate ACK is to let the sender know that a segment was received out of order, and also, the sequence number expected.
- 3. However, if a segment at the end of a sequence of segments is lost, there are no subsequent segments to generate duplicate ACKs. In this case, the sender's *retransmission time* (discussed below) will expire (a *timeout* will occur) since there are no corresponding ACKs for this segment and the sender cannot wait indefinitely for a delayed ACK.

From cases 1 and 2, it is difficult for a TCP sender to know whether the reception of a duplicate ACK is an indication of a lost segment or just an out-of-order data delivery. Thus, a TCP sender must wait for a small number of duplicate ACKs (typically three) before deciding whether a segment is lost or simply out of order. The assumption here is that if there is just an out-of-order delivery, the receiver will issue only one or two duplicate ACKs before the out-of-order segment is processed, at which time a new ACK will be generated. However, if three or more duplicate ACKs are issued in a row, then there is a strong chance that a segment has been lost.

When the TCP sender receives three duplicate ACKs, it immediately retransmits the lost segment (fast retransmit) and enters into a phase called fast recovery. Fast retransmit enables a TCP sender to rapidly recover from a single lost segment, or one that is delivered out of sequence, without shutting down the *cwnd*. The receiver responds with a cumulative ACK for all segments received up to that point. *Fast recovery* is based on the notion that, since the subsequent packets generating the duplicate ACKs were successfully transmitted through the network, there is no need to enter slow start and dramatically reduce the data transfer rate; therefore cwnd should stay open. Thus, in fast recovery, the ssthresh is set to half of the current window size and cwnd is set three segments greater than ssthresh to allow for three segments already buffered at the receiver. On the arrival of each additional duplicate ACK, cwnd is incremented by a segment allowing more data to be sent. When an ACK arrives that acknowledges new data, *cwnd* is set back to ssthresh, and TCP enters the congestion avoidance mode.

In the congestion avoidance mode, the TCP sender increments *cwnd* by one segment every RTT until data loss occurs or the receiver's advertised window is reached. If the loss is isolated to a single segment, the resultant duplicate ACKs will cause the sender to halve *cwnd* and then continue a linear growth of *cwnd* from this new point.

The expiration of a retransmission timer may indicate more serious congestion. In this case, the *ssthresh* is set to half of the current window size. But because the sender has not received any useful information for more than one RTT, it adopts a more conservative approach where it closes the congestion window *cwnd* back to one segment, and resumes data transmission in the slow-start mode.

Experimentation has shown that starting off with a larger initial window size of three or four segments in slow start will allow more segments to flow into the network, generating more ACKs, and will decrease the time it takes to complete slow start [12]. Because the *cwnd* can open up faster as a result, better performance is gained, in particular for small files transmitted over links with long RTTs. The short-duration TCP sessions typical of Web fetches can benefit a lot from this feature. However, a starting value of four segments in slow start may be too many for low-speed links with limited buffers, so a starting value of not more than two is a more robust approach [12].

4.2.3. TCP Timeout and Retransmission. To provide reliable data transfer, a TCP receiver is expected to acknowledge the data that it receives from the sender. Every time a segment is sent to a receiver, the sending TCP starts a timer and waits for an acknowledgment. If the timer expires before the receiver acknowledges receipt of the data in the segment, the sending TCP assumes that the segment was lost or corrupted and retransmits it.

TCP monitors the delay performance of each connection to derive reasonable estimates for timeouts. The timeout

estimates are adjusted from time to time to account for changes in the delay performance of the connection. This is because successive TCP segments in a connection may not be sent on the same path and may experience different delays as they traverse different sets of routers and links. To accommodate the varying network delays experienced by segments in a connection, TCP uses an adaptive retransmission algorithm. The measurement of the round-trip time (RTT) experienced on a connection forms the key input to this retransmission algorithm.

TCP measures the elapse time between sending a data byte with a particular sequence number and receiving an acknowledgment that covers that sequence number. This measured elapse time is the sample round-trip time (*Sample_RTT*). On the basis of the *Sample_RTT*, a smoothed round-trip time (*SRTT*) is computed using the exponentially weighted moving average filter

$$SRTT \leftarrow \alpha \times SRTT + (1 - \alpha) \times Sample_RTT$$

where α is a weighting factor whose value is between 0 and 1 ($0 \le \alpha < 1$). TCP then computes a retransmission timeout (RTO) value that is a function of the current estimated round-trip time (*SRTT*). RFC 793 [1] recommended the RTO to be computed as

$RTO = \beta \times SRTT$

where β is a constant delay weighting factor ($\beta > 1$) with a recommended value of 2. It has been shown [9] that this computation does not adapt to wide fluctuations in the RTT, thus causing unnecessary retransmissions that add to network traffic. Jacobson [9] suggests that both the smoothed RTT estimates (i.e., mean values) and the variance in the RTT measurements should be maintained. These can then be used to compute the RTO instead of just computing the RTO as constant multiple of the smoothed RTT. These changes make TCP more responsive to wide fluctuations in the round-trip times and yield higher throughput. Thus, the following computations are required for each RTT measurement:

$$\begin{array}{l} Diff = Sample_RTT - SRTT\\ SRTT \leftarrow SRTT + \eta \cdot Diff\\ Dev \leftarrow Dev + \theta \cdot (|Diff| - Dev)\\ RTO = SRTT + \phi \cdot Dev \end{array}$$

where *Dev* is the smoothed mean deviation, η and θ are filter gains with values in the range $0 < \eta, \theta < 1$, and ϕ is a factor that determines how much influence the deviation has on the RTO. For efficient TCP implementation, η and θ are typically selected to each be an inverse power of 2 so that the operations can be done using shifts instead of multiplies and divides. Jacobson [9] specified $\phi = 2$, but after further research, it was changed to $\phi = 4$ [13].

Although the RTO estimation process described above is relatively straightforward, a problem occurs when a segment is retransmitted. Let us say that TCP transmits a segment, a timer expires and TCP retransmits the segment. When an acknowledgment is received, the sender has no way of knowing whether the acknowledgment corresponds to the original or retransmitted segment. Perhaps the first segment was delayed and not lost or the ACK for the first segment was simply delayed. This situation is sometimes referred to as acknowledgment ambiguity. Now, what should TCP do with regard to RTT estimates when the TCP acknowledgments are ambiguous? The answer to this problem is provided by Karn's algorithm [14], which adjusts the estimated round-trip time only for unambiguous acknowledgments. When retransmissions take place in Karn's algorithm, the SRTT is not adjusted. Instead, a timer backoff strategy is used to adjust the timeout by a factor. The idea is that if a retransmission occurs, something drastic could have happened in the network, and the timeout should be increased sharply to avoid further retransmission. The timer backoff strategy is implemented with a multiplicative factor γ as

$$New_timeout = \gamma \times timeout$$

Typically, γ is set to 2 so that this algorithm behaves like a binary exponential backoff algorithm. The algorithm uses the SRTT to compute an initial timeout value (*timeout*), then backs off the timeout on each retransmission (to obtain New_timeout) until a segment is successfully transmitted. The timeout value that results from backoff is retained when subsequent segments are sent. When an acknowledgment is received for a segment that does not require retransmission, TCP measures the RTT value (Sample_RTT) and uses this value to compute the SRTT and the RTO values.

5. TCP FUTURES

The rapid growth of the Internet continues to evolve TCP in several ways. In particular, it is becoming more apparent that enhancements to TCP are required to address the challenges presented by the different transmission media encountered in the Internet. Highspeed (optical fiber) links, long and variable delay (satellite) links, lossy (wireless) networks, asymmetric paths (in hybrid satellite networks), and other linkages are becoming widely embedded in the Internet. These different transmission media present a wide range of characteristics that can cause degradation of TCP performance. For example, long propagation delay and losses on a satellite link, handover and fading in a wireless network, bandwidth asymmetry in some media, and other phenomena have been shown to seriously affect the throughput of a TCP connection. Also, TCP assumption that all losses are due to congestion becomes quite problematic over wireless links. TCP considers the loss of packets as a signal of network congestion and reduces its window consequently. This results in severe throughput deterioration when packets are lost for reasons other than congestion. Noncongestion losses are mostly caused by transmission errors.

Some of the techniques currently under study to enhance TCP performance involve modifying TCP to help it cope with these new media types. Others keep the protocol unchanged, but place some intelligence in the intermediate network elements along TCP connections to provide faster and more accurate feedback to the TCP endpoints about the conditions on the connection.

BIOGRAPHY

James Aweya received his B.Sc. degree in electrical and electronics engineering from the University of Science & Technology, Kumasi, Ghana, the M.Sc. degree in electrical engineering from the University of Saskatchewan, Saskatoon, Canada, and a Ph.D. degree in electrical engineering from the University of Ottawa, Canada. Since 1996, he has been with Nortel Networks where he is currently a systems architect in the Advanced Technology Group. His current activities include the design of resource management and control functions for communication networks, the development and analysis of new network architectures and protocols, and the design and analysis of switch and router architectures. He has published more than 50 journal and conference papers and has a number of patents pending. Dr. Aweya was the recipient of the IEEE Communications Magazine Best Paper Award at OPNET-WORK 2001. In addition to communication networks and protocols, he has other interests in neural networks, fuzzy logic control, and application of artificial intelligence to computer networking. Dr. Aweya also collaborates with a number of researchers at the University of Ottawa and Carleton University, Ottawa, Canada, to research issues on communication network design and quality of service control.

BIBLIOGRAPHY

- 1. J. Postel, *Transmission Control Protocol*, RFC 793, Sept. 1981.
- 2. J. Postel, User Datagram Protocol, RFC 768, Aug. 1980.
- S. Iren, P. D. Amer, and P. T. Conrad, The transport layer: tutorial and survey, ACM Comput. Surv. 31(4): 360-405 (Dec. 1999).
- 4. V. Jacobson, R. Braden, and D. Borman, *TCP extensions for High Performance*, RFC 1323, May 1992.
- M. Mathis, J. Madavi, S. Floyd, and A. Romanov, TCP Selective Acknowledgment Options, RFC 2018, Oct. 1996.
- J. Mogul and S. Deering, Path MTU Discovery, RFC 1191, Nov. 1990.
- R. Braden, ed., Requirements for Internet Hosts Communication Layers, RFC 1122, Oct. 1989.
- J. Nagle, Congestion Control in IP/TCP Internetworks, RFC 896, Jan. 1984.
- V. Jacobson, Congestion avoidance and control, ACM Comput. Commun. Rev. 18(4): 314–329 (Aug. 1988).
- W. R. Stevens, *TCP/IP Illustrated*, Vol. 1, Addison-Wesley, Reading, MA, 1994.
- M. Allman, V. Paxson, and W. Stevens, TCP Congestion Control, RFC 2581, April 1999.
- M. Allman, S. Floyd, and C. Partridge, *Increasing TCP's Initial Window*, RFC 2414, Sept. 1998.
- V. Jacobson, Berkeley TCP evolution from 4.3-Tahoe to 4.3 Reno, Proc. 18th Internet Engineering Task Force, Univ. British Columbia, Vancouver, BC, Sept. 1990.

 P. Karn and C. Partridge, Improving round-trip time estimates in reliable transport protocols, *Comput. Commun. Rev.* 17(5): 2-7 (Aug. 1987).

TRANSPORT PROTOCOLS FOR OPTICAL NETWORKS

TIM MOORS University of New South Wales Sydney, Australia

1. INTRODUCTION

Transport protocols are used in the end systems that are connected by communication networks to match the characteristics and requirements of the network to those of the users of the communication network in the end systems. In particular, transport protocols are used to match the aspects of reliability, security, transmission unit size, and timing of information transfer. For example, when the network can corrupt, lose, missequence, or duplicate information, a transport protocol may be used to enhance the reliability of the end-to-end path to match the requirements of end users. A transport protocol may also segment information from its users to meet the maximum transmission unit requirements of the network, and may pace the transmission of information supplied by the user so as to prevent or limit congestion in the network. In addition to matching the network path to users, a transport protocol may also be used to match users to each other, such as by controlling the flow of information from source to destination(s) so that its rate does not exceed the capacity of the destination(s).

Optical networks have grown in importance since the introduction of low-attenuation fiber in the 1970s. This article focuses on networks constructed from optical fiber, although many of the principles also apply to networks that employ free-space optics. The first generation of optical fiber networks used optical transmission systems, and converted the optical signal to electronic form for switching. More recently, optical networks have appeared in which the switching is also done optically, creating "all-optical networks," in which a "lightpath" extends between communicating endpoints, with optoelectronic conversion only occurring in the end systems, not along the end-to-end path.

Since the end systems, where transport protocols are implemented, always provide optoelectronic conversion (even when these end systems are connected to all-optical networks), transport protocols are generally implemented in the electronic domain. This leads to a common misconception that existing transport protocols that were designed for electronic networks are also appropriate and optimal for optical networks, and directs the attention for innovation in optical networks toward the transmission and switching systems. In reviewing transport protocols for optical networks, this article will show the fallacy of this misconception: that transport protocols are *strongly* affected by the existence of optical networks.

2614 TRANSPORT PROTOCOLS FOR OPTICAL NETWORKS

This article starts by reviewing the features of optical networks that are particularly salient to transport protocols. It then provides an overview of the generic functions that transport protocols are expected to provide, and discusses how these functions are affected by the use of optical networks. Section 4 then describes specific transport protocols that have been used with optical networks. Finally, transport protocols for optical networks need to be implemented in a manner that is appropriate for the high speeds of optical networks. In Section 5, this article concludes by considering such implementation issues.

Before considering transport protocols for optical networks in depth, it is worth making two points about terminology and the scope of transport protocols:

- 1. The term "transport protocol," as used in this article and in the context of computer communication systems [e.g., 1], refers to an end-to-end protocol. This can be confusing because in the context of optical transmission systems (in particular, in the context of the International Telecommunication Union's G series of recommendations [e.g., 2]) the term "transport" often refers to the transmission of information between adjacent points, such as across one of the multiple hops of an end-to-end communication path.
- 2. While the most prominent role of transport protocols in optical networks is in end systems, they are also used in routers, switches, and other intermediate systems to ensure the reliable delivery of information between such systems. For example, the popular Border Gateway Protocol for routing sends routing topology update information over the transport protocol TCP.

Despite these different uses of the term "transport" and different applications of transport protocols, this article will focus on the end-to-end transport protocols that are used with optical networks.

2. SALIENT FEATURES OF OPTICAL NETWORKS

Optical fibers can be readily manufactured with loss of a couple of decibels (e.g., 1-4 dB) per kilometer. This allows optical networks to span vast distances, and a transport protocol for an optical network may need to deal with large propagation delays, simply because of the large distances and the limited propagation speed of light. Comparing networks that span the same distance, optical networks may exhibit lower end-to-end delay than their electronic or radio-based counterparts. This is not due to inherently faster propagation of optical signals, since they propagate at a speed of around 2×10^8 m/s in silica fiber-a figure that is comparable to the propagation speed in electrical wiring or radio transmission. Rather, optical networks tend to reduce the delays that the signal will incur as it passes through intermediate systems for two reasons: (1) optical networks tend to minimize buffers because they are expensive-it is difficult to buffer information in the optical domain (e.g., using delay lines built using long threads of fiber) for alloptical networks, and electronic buffers that can match the bandwidth of optical transmission systems are expensive for optoelectronic networks; and (2) the high transmission rates of optical networks mean that large volumes of information can be "in flight" from source to destination at any time. Rather than allow all this information to accumulate in a buffer in a router or switch (which would require inordinately large buffers), optical networks tend to be used with connection-oriented call admission and congestion control schemes that limit the burstiness of traffic entering points of the network, and so reduce the delays incurred by buffering.

Optical fibers exhibit low loss for wavelengths in the range of 1280-1620 nm. This broad range of wavelengths creates a bandwidth of ~ 30 THz, so with proper modulation, each fiber has the capacity to carry information at rates of terabits per second. At present, lineterminating equipment (e.g., lasers and receivers) can only operate at Gbps (gigabits per second) rates, so the capacity of the fiber can be exploited only by multiplexing multiple signals on a fiber, such as by using wavelength-division multiplexing (WDM). Even so, the transmission rate of optical networks is vastly higher than that of electronic or radio systems, and optical networking products are currently even available for the consumer market with transmission rates of 1 Gbps. The high transmission speed of optical networks shifts the performance bottleneck from transmission links to processing and buffering systems. The high transmission speed also reduces the effect of the transmission time on the end-to-end delay, and, instead, the end-to-end delay becomes dominated by propagation delays that are essentially fixed by the speed of light [3]. The only way to reduce the delay for such transmissions is to reduce the distance that signals must propagate, for example, by reducing the number of round-trip times involved in signaling, or by retrieving information from caches that are near the destination.

Optical transmission systems exhibit extremely low bit error rates compared to electrical (e.g., coaxial cable) or wireless transmission systems. For example, optical transmission systems often exhibit bit error rates of the order of 1×10^{-11} [4], and have been reported with error rates as low as 1×10^{-15} . While the line error rates may be low, random transmission errors may still occur on buses internal to routers and switches. In optical networks, there may also be burst loss or errors caused by buffer overflow, or by fixed-duration events (e.g., protection switching) affecting large numbers of bits. Thus, while optical transmission error rates may be low, transport protocols still need mechanisms to ensure reliable transfer.

The security of optical fiber systems is often considered to be stronger than that of other transmission media because the fiber confines the propagation of the optical signal. While it is still possible to tap fiber in an almost undetectable manner, this tends to be more difficult than intercepting electrical or wireless communications.

Compared to electronic networks, optical networks shift the emphasis from packet switching toward circuit switching. This is often done because either optical

processing within the network is difficult (e.g., WDM for all-optical networks), to reduce buffering within the network, or in order to simplify electronic processing to match it to the rate of optical links [e.g., label swapping technologies such as multiprotocol label swapping (MPLS) and asynchronous transfer mode (ATM)]. In the case of WDM, the circuits may be real lightpaths, while in the cases of MPLS/ATM, the circuits tend to be "virtual," in that they may define a path from source to destination, but may not reserve resources for the exclusive use of traffic that uses the "virtual circuit." Circuits can isolate the traffic of one source from that of others, making transportlayer congestion control redundant. With some optical networks [e.g., those based on WDM or the synchronous digital hierarchy (SDH)], the granularity of bandwidth that can be assigned to a virtual circuit may be coarse (e.g., in units of 51 Mbps for SDH), and this elevates the importance of multiplexing by the transport protocol. Circuits also preserve the sequence of information that they carry, simplifying the transport protocol function of error control. (On the other hand, in some optical networks, the difficulty of buffering in the optical domain leads to switches "deflecting" incoming information that is destined for an outgoing port that is busy. Such deflection routing [5] can lead to significant missequencing.)

When an optical lightpath extends end-to-end between communicating terminals, the optoelectronic conversion is physically collocated near the implementation of the transport protocol. This specialized hardware can include hardware support for transport protocols for optical networks, including calculation of integrity checks.

3. TRANSPORT PROTOCOL FUNCTIONS

The introduction to this article stated that transport protocols "match the characteristics and requirements of the network to those of the users of the communication network in the end systems ... [and] may also be used to match users to each other." This section describes, in depth, the functions that a transport protocol may implement to provide this matching, and how they are affected by the presence of optical networks. The functions covered are as follows: reliable transfer (Section 3.1); flow and congestion control (Section 3.2); security (Section 3.3); framing, segmentation, and reassembly (Section 3.4); multiplexing (Section 3.5); and state management (Section 3.6).

The scope of this section (and, indeed, this article) is limited to *unicast* transfers from a single source to a single destination. Multicast is another important mode of communication; however, multicast transport protocols that provide functions such as reliability are still in their infancy, and the impact of optical networks on these transport protocols is still uncertain. Furthermore, while this section considers the impact of traffic that has stringent timing requirements (e.g., voice and video) on the function of multiplexing, it does not address how transport protocols (such as the Real-time Time Protocol [6]) may provide functions that reconstruct the timing of such traffic, since this function is independent of the existence of optical networks.

3.1. Reliable Transfer

Many applications seek assurance of the integrity of the information that they exchange. Transport protocols provide this assurance by adding well-defined redundancy (e.g., cyclic redundancy codes or checksums) to the payload information that they send, and destinations check that the information that they receive preserves this redundancy, suggesting that it has retained its integrity. If a discrepancy is found, then errors may be corrected either using the redundancy (forward error correction) or the source retransmitting the information. To limit the volume of information that a source must retransmit, and so improve transmission efficiency, it is common for the source to segment the information that it transmits into smaller parts (segments) that are often carried in separate packets that flow through the network.

There are six aspects to the reliable transfer of these segments, which can be understood in analogy to transferring the sections (segments) of this article from the author to the reader:

- 1. *Integrity*. Received segments should have the same content as transmitted segments; for instance, no typographic errors should be introduced.
- 2. *Uniqueness*. Each segment sent should be received only once; for instance, sections of the article should not be duplicated.
- 3. *Completeness*. Each segment sent should be received; for instance, no sections should be missing from the midst or the end of the article.
- 4. *Sequence*. Each segment should be received in the proper position relative to other segments.
- 5. *Relevance*. No extraneous segments (e.g., from other sources) should be inserted in the midst of the segments sent by the source.
- 6. *Delivery*. In addition to these five aspects of reliability that may interest the recipient, there is another aspect that may interest the source. The source may be interested in whether the destination successfully received all segments.

Dividing the broad concept of "reliability" into six aspects is important because different applications are concerned with different aspects of reliability, and these concerns translate into the functionality that these applications seek of the transport layer. For example, real-time media such as voice and video tend to be less concerned with integrity and completeness than they are concerned with sequence. Transaction systems are often unconcerned with delivery acknowledgments, because the client (who is the source of the request, and destination of the response) will issue another request if a response is not received.

Optical networks have differing effects on how transport protocols address the six aspects of reliability. The high *integrity* of optical networks allows transport protocols to use larger segments than for other networks, while still retaining the same transmission efficiency. Larger segments are important because they help reduce the rate at which segments (and packets) must be processed, helping redress the difference in processing and transmission rates introduced by optical networks. The integrity of optical transmission systems also suggests that transport protocol integrity checks should emphasize detection of errors introduced in switches and routers, rather than those introduced on the transmission line. The performance of integrity checks is highly sensitive to implementation techniques that are discussed in Section 5 of this article.

While optical networks generally preserve packet sequence, transport protocols still tend to use sequence numbers to identify segments that need to be retransmitted. The preservation of sequence also helps destinations detect in complete delivery. Consecutive received segments that do not have consecutive sequence numbers suggest that either intermediate segments have been lost, or the latter segment is a retransmission-a case that can be identified if retransmitted segments are explicitly identified as such in their headers. This allows the retransmission process to be initiated by the destination sending a negative acknowledgment to the source when it detects missing information, rather than the source having to time out while waiting for an acknowledgment from the destination. This allows prompt retransmission, rather than waiting for a loosely calibrated timeout, and eliminates the need for complicated timers based on round-trip time measurements [7]. Such negative acknowledgments are used by several transport protocols designed for highspeed optical networks, including XTP [8], NETBLT [9], VMTP [10], and SSCOP [11].

Optical networks can help the transport protocol to estimate the round-trip time, if positive acknowledgments *are* used, or if this is needed for congestion control purposes. This is because optical networks have limited buffering and may provide rate guarantees, thus reducing the variation in round-trip times, and allowing timeouts to be calibrated more accurately. The small buffers also reduce the maximum packet lifetime in the network, which reduces the period for which the transport protocol needs to preserve state after the active data transfer has ended in order to ensure that only *relevant* segments are delivered. In TCP, this period is called the "time wait" period, and reducing it reduces the volume of state information that an end system must maintain, which is important for busy servers.

The high bandwidth of optical transmission systems can lead to a large volume of information being "in flight" between source and destination at any time. Transport protocols must choose between using selective retransmissions and a "go-back-n" policy in which the source retransmits everything from the damaged segment onward, including segments that were originally received intact. If errors are truly rare, then the transmission efficiency of go-back-n may be tolerable; otherwise the transport protocol should use selective retransmissions, which require a resequencing buffer in the destination whose size increases with the transmission rate. Like the resequencing buffer at the destination, sources need to maintain a retransmission buffer that may be large for optical networks since its size increases with the volume of information "in flight." Finally, transport protocols for optical networks need to have large-sequence-number fields so that they can uniquely identify each segment (or byte) of information that is in flight from source to destination. These large sequence number fields are used for both reliable transfer and for flow control.

3.2. Flow and Congestion Control

Flow control ensures that a source does not transmit at either a sustained rate or in bursts that exceed the capacity of the destination to process incoming traffic. Flow control works by the source and destination agreeing on a constraint that will govern the source's transmissions, and the destination sending feedback to update this constraint. Traditional transport protocols have constrained the volume of information sent; the protocol allows the source to transmit a certain number of bytes or segments before it must wait to receive feedback from the destination, allowing the source to transmit more. The high bandwidth of optical networks, combined with even modest transmission delays, can give the end-toend path a large "bandwidth-delay product," meaning that the source must have a large volume of information outstanding at any time in order to transmit at the rate of the network. Unless the source spreads the transmission of this large volume of information over some interval, it will transmit very bursty traffic that is likely to lead to congestion in network elements.

One way to address avoid bursts is to use the self-clocking "slow start" [12] technique, in which the source transmits a small window of information, and then expands this window (up to some limit) with each acknowledgment that it receives from the destination. While this technique avoids an initial burst, it prevents the source from transmitting at the full rate until some interval after the initial transmission. This significantly affects sources that have small amounts of information to transmit, since they may always have their transmission rate limited by this slow-start regime.

An alternative approach to flow control that is popular amongst transport protocols for optical networks [e.g., 8-10] is to use "rate"-based control, where the constraint on source transmissions is defined in terms of a transmission rate rather than a window. Often the "rate" is specified in terms of bits per second, or a mean segment interarrival time, *in addition to* a burst size, which limits the volume of information that the source can send at any instant.

Rate-based controls are complicated by the fact that few current end systems employ real-time operating systems that can guarantee to a process the resources needed to sustain consumption of traffic at a certain rate, independent of competing demands from other processes. Furthermore, few processes can predict how much processing incoming information will require. Thus, while a destination process or transport protocol may propose a rate that it can accept, it will often monitor the level to which its buffers are filled in order to decide whether to adjust the rate up or down. Rate-based controls can also be complicated by operating systems that provide timers that have coarse resolution, since maintaining a high transmission rate with such timers requires that
large bursts also be possible. On the positive side, ratebased controls are becoming increasingly common at the interface between the transport and network layer of modern optical networks. Here, rate-based controls are used to "condition" (also known as "shape" and "police") traffic entering the network so that it doesn't overwhelm the network; either overflowing the limited buffers in an optical network, or interfere with traffic from other sources that has tight delay constraints. Some transport protocols, such as XTP [8], allow sharing of rate control information between the network and transport layer, so that the rate control by the transport layer can also help satisfy network rate requirements.

While flow control prevents a source from overloading the limited capacity of a *destination* endpoint, congestion control is designed to prevent a source from overloading the network's capacity. While the transport layer may know how the application would prefer to respond to congestion (e.g., by slowing down, or discarding, traffic), it is the network layer that is ultimately responsible for ensuring that this response occurs, preventing an end system from sending so much traffic that the network congests, to the detriment of other network users. In the past, it was often convenient to implement congestion control in the transport layer (e.g., congestion control was added to the Internet through a small change to TCP [12]). However, networks cannot rely on transport layer congestion control because transport protocols are implemented in end systems where users, who may prefer to place their interests before those of the network, can readily change them. Transport layer congestion control may also be inappropriate for optical networks that provide end-to-end circuits, since the congestion control may only unnecessarily delay source transmissions.

3.3. Security

Although fibers give optical signals some physical security, providing logical security requires encryption of the information transmitted over the optical network. According to the end-to-end arguments [13], the transport protocol is a natural location for encryption, since it is implemented in the end systems, where users can control and trust the encryption and keys that are used. However, encryption of data for optical networks is challenging because of the computational complexity of cryptographic functions, when processing is already a bottleneck compared to the transmission system. This encourages hardware implementations of cryptographic functions [14].

Transport protocols for optical networks need to consider the dependency of many encryption systems on processing data in sequence. For example, traditional encryption often uses ciphers in closed-loop cipher feedback or cipher block chaining modes, in which the output from encrypting initial data is fed back into the cipher to influence the encryption of subsequent data. Such serial processing restricts the possibility of parallelism in the encryption process, which could raise the throughput to match optical transmission rates. Consequently, optical networks may favor ciphers used in an output feedback mode, in which multiple cryptographic systems operate in parallel to generate streams of bits that are unpredictable without the correct key, and the payload is simply (and rapidly) exclusive-ored with these bitstreams to protect it prior to transmission.

While a transport protocol may be expected to enhance security through encryption of the payload information, it should not itself introduce security vulnerabilities of its own. For example, transport protocols introduce state information, which can lead to denial of service attacks (whereby an attacker attempts to exhaust all state storage space available on a server, preventing service to genuine clients). Such attacks can be countered by the use of cryptographic "cookies" [15]. Similarly, while optical networks may promote the use of negative acknowledgments for error control, these should include information that allows the source to authenticate that they came from the destination. Otherwise, a simple denial of service attack can be effected by a third party repeatedly sending negative acknowledgments to the source, forcing it to retransmit information rather than transmitting new information

3.4. Framing, Segmentation, and Reassembly

Packet-switched networks impose limitations on the maximum length of each packet for reasons such as controlling the delay that one packet may experience when serialized behind another packet in a multiplexer. A transport protocol for an optical network may segment information from the application prior to transmission to satisfy these maximum transmission unit requirements of the network, and later reassemble it at the destination. It may also use segmentation and reassembly to limit the size of information that needs to be retransmitted when recovering from a bit error, and this is done for both packetand circuit-switched networks. Whenever segmentation is performed, it may be important for the optical network to be aware of where transport layer segments begin and end so that it can discard a few whole segments, rather than many segment parts, using techniques to be described in Section 4.2.

While the network may provide a conduit that carries information from source to destination, the data units that applications send across this conduit are often discrete, and should not be merged. Several transport protocols provide framing of application data units, so that the destination can determine where, among the information received, application data units start and end. Some transport protocols (notably TCP) provide a bytestream only between communicating applications, and applications can indicate where their data units begin and end only by creating separate connections for each data unit.

3.5. Multiplexing

While the network delivers information to the appropriate *network interface*, often identifying a physical entity such as a computer, the transport layer is responsible for delivering this information to the appropriate *process* operating in the device that has that network interface. Thus, transport protocols provide multiplexing at sources to allow multiple processes to share a single network

interface, and demultiplexing at destinations to direct incoming information to the appropriate interface.

Multiplexing tends to be implemented by including in transport layer segments fields that indicate the destination and source ports of the segment, although transport layers may also use the multiplexing identifiers that are used by connection-oriented network layers (such as ATM) in order to avoid the harmful effects of layered multiplexing [16]. Since some optical networks offer circuits only with coarse granularity of bandwidth, multiplexing is particularly important to allow multiple processes on the end systems to share the large end-toend bandwidth. When multiplexing together traffic from multiple applications, the traffic from one application (e.g., a delay-insensitive file transfer) may interfere with the delay experienced by other applications (e.g., a delay-sensitive telephone conversation). Some more recent transport protocols designed to support multimedia [e.g., 17] include schedulers that are designed to avoid such interference.

3.6. State Management

Unlike the functions discussed in preceding sections that directly satisfy network or application needs, transport protocols use state information to support their implementation of other functions. In particular, state information is important in the provision of reliable transfer, flow, and congestion control. State information can also be used to record negotiated values of parameters that govern the operation of the protocol during the transfer, including segment size or whether end systems should represent sequence numbers in big- or littleendian format [18]. Such negotiation is often relatively simple for transport protocols since for unicast transfer, they involve only two parties that must negotiate. Such negotiation is important for transport protocols for optical networks since end systems that can agree to simplify the communication may be able to employ simpler protocols, and so keep up with the rate of optical transmission systems. An extreme form of negotiation is where transport protocols are composed [19,20] on demand to include minimal functionality that satisfies source, destination, and network requirements. Such composition has the potential to eliminate unnecessary functionality, and so simplify transport protocols so that they can match the transmission rates of optical networks.

The signaling to establish or release state information can be done either *in band*, on the same channel that is used for data transfer, or *out of band*, on a separate channel. Sending signaling information on the same channel that is used for data transfer avoids the potential problem of completing signaling when no data channel is available, but has the disadvantage of complicating the processing that is required on the frequently used data channel with rarely used and complicated signaling functionality. Most transport protocols to date (e.g., TCP) have emphasized in-band signaling, but out-of-band signaling is used in the Advanced Peer-to-Peer Networking protocol [21] and may become more widespread with alloptical networks in which network signaling must be performed electronically, creating an electronic control plane that is separate from the optical data plane. ATM also emphasizes separation of the control and "user" (data) planes.

Servers that serve multitudes of clients (e.g., server of a popular Website) often require the high transmission rate of optical networks. For such servers, the overhead of retaining state information for each client for long periods can become onerous, and this promotes the conveyance of state information through cookies [15]. Note that it is the number of clients that causes this issue, and leads to the server using a high-speed (e.g., optical) network. That is, this issue is *correlated* to the use of optical networks, but is not *caused* by the use optical networks.

4. SPECIFIC TRANSPORT PROTOCOLS FOR OPTICAL NETWORKS

This section describes specific transport protocols for optical networks, namely, TCP, those relating to ATM, and XTP. It emphasizes the match between the functionality of these protocols and that required for optical networks. While the functionality of a protocol is important, the method by which that functionality is implemented also affects the suitability of a transport protocol to optical networks, and will be addressed in Section 5.

4.1. Transmission Control Protocol

George Santayana wrote, "Those who cannot remember the past are condemned to repeat it" and this has led to a networking proverb: "Those who ignore TCP are doomed to reinvent it." The Transmission Control Protocol (TCP) is currently the most widely used transport protocol on existing electronic and optoelectronic networks. While originally introduced in the 1970s [22], TCP has since been extended with new options [e.g., 23] that improve its suitability to high-speed optical networks. This section considers the application of TCP to optical networks. The key features of TCP are its provision of all six aspects of reliability, and the minimal information that it needs from the network layer, which allows it to operate over varied networks, including optical networks.

TCP provides reliable transfer by including in the header of each segment a 32-bit sequence number field, a 32-bit acknowledgment field, and a 16-bit checksum field. During connection establishment, the source and destination agree on the initial sequence number to use for the transfer, and the source sets the sequence number field in each segment that it sends to indicate the position of the first byte of the segment in the sequence number space. For a TCP destination to correctly locate incoming segments, it must ensure that the sequence number of each segment is unique relative to other segments that are also in transit from the source. For a 1-Gbps transmission system, this means that each segment should take no longer than 17 s to propagate through the network [23]. A timestamp option allows TCP to operate across networks with higher maximum segment lifetimes [23].

A TCP destination uses the checksum field to verify the integrity of incoming segments. Whenever the destination receives two maximum-sized segments worth of data (or

500 ms elapses since receiving a segment that it has not yet acknowledged), it sends a cumulative positive acknowledgment to the source to indicate the sequence number of the next segment that it is expecting to receive, that is, that follows any contiguous set of segments that it has received so far. If a segment is lost, then the destination will send duplicate acknowledgments when it receives subsequent segments, acknowledging receipt of the same contiguous set of segments. The source waits a certain period after transmitting a segment for that segment to be acknowledged, and if an acknowledgment is not forthcoming, then it will retransmit the segment. This timeout may occur after a reasonably long time, and with high-speed optical links, the source may stall before this time, being forced by flow or congestion control mechanisms to defer additional transmissions until it receives a new acknowledgment. The "fast retransmit" extension to TCP [24] can circumvent this stalling on high-speed links, by the source retransmitting a segment when it receives multiple duplicate acknowledgments that suggest loss of that segment, rather than waiting for a timeout to occur. Selective acknowledgments have also been added as a TCP option [25] to improve transmission efficiency on high-speed long-delay paths.

The congestion control of TCP [12] is related to its error control through the interpretation of suggestions of segment loss as indicators of congestion. This interpretation is more valid for optical networks than it is for wireless networks, where appreciable loss can also occur as a result of transmission errors. More recently, TCP has been extended to allow processing of explicit congestion notifications from routers [26]. A TCP source will gradually increase its transmission rate, by one segment per roundtrip time, while it does not observe signs of congestion (duplicate acknowledgments, retransmission timeouts, or explicit congestion notifications). When a TCP source observes signs of congestion, it quickly halves its transmission rate. This additive-increase, multiplicative-decrease behavior allows a TCP source to adapt to changing network conditions, and helps multiple sources converge on a "fair" allocation of bandwidth. The fact that a TCP source discovers the permissible transmission rate by itself makes this congestion scheme remarkably general, being able to operate over wireless and optical packet-switched and circuit-switched networks, as well as being flexible and thus able to adapt to changing network conditions. However, for optical networks, a TCP source can take some time to ramp up its transmission rate to that allowed by the network, even when using the slow-start function. Furthermore, a TCP source will continue to probe for additional bandwidth, increasing its transmission rate until segments are lost (e.g., due to source transmission buffers overflowing) and then back off. This behavior is suboptimal for optical networks in which the capacity of a (virtual) circuit may be fixed—increasing the rate only leads to unnecessary loss, and backing off unnecessarily slows the transmission. While there has been research into improving TCP when the network can provide minimum rate guarantees [27], further research is needed to optimize the performance of TCP on optical networks so

that the source transmission rate converges on the path capacity, rather than oscillating around it.

The flow control mechanism of TCP originally limited its performance on high-speed optical networks. TCP's flow control works by the destination setting a field in segments returning to the source to indicate the size of its receive window. The 16-bit size of this field, and the fact that it measured bytes, limited a TCP source to having at most 65536 bytes of information unacknowledged at any time — a volume that is insufficient to fill many high-speed transmission links. This has since been corrected by the addition of a TCP window-scale option [23] that effectively increases the size of the receive window to 32 bits.

While TCP has successfully evolved to match the capacity and requirements of optical networks, it is showing signs of its age through the susceptibility of its state management to denial of service attacks, and lack of support for multihoming and partially ordered delivery. While none of these perceived deficiencies particularly limit its applicability to optical networks, it is likely that TCP will be succeeded in the future by newer protocols such as the Stream Control Transmission Protocol [15].

4.2. The Impact of ATM

The asynchronous transfer mode (ATM) was heralded in the early 1990s as a technology that would replace existing telephony and packet-switched technologies (e.g., Ethernet and Frame Relay) with a single broadband integrated services digital network. It was designed with optical networks in mind, by fixing the packet ("cell") size and including in each cell labels ("virtual channel" and "path" identifiers) that were intended to facilitate high-speed switching that could match the rate of optical transmission systems. ATM affects transport protocols in two ways. First, ATM adaptation layers were designed to adapt the native ATM service into a service that better matches the requirements of communicating applications, essentially creating a set of new transport protocols. The Service Specific Connection Oriented Protocol (SSCOP), described below, is an example of one of these transport protocols. The second effect results from the different framing used by ATM and transport protocols, and led to particular discard techniques within the optical network to accommodate transport protocols.

The impetus for SSCOP's design [11] was high-speed user data transfer, but it was first used to carry signaling in ATM networks. SSCOP exploits the fact that ATM networks preserve sequence (as do many optical networks), allowing the destination to detect loss when it receives a segment with a sequence number that does not follow its predecessor. An SSCOP destination will then immediately send an unsolicited status message (negative acknowledgment) to the source that requests retransmission of the missing segment. SSCOP is simple, allowing it to keep up with optical transmission speeds, by virtue of only requiring one timer at the source to trigger the transmission of periodic poll messages to the destination. SSCOP recovers from segment loss by the destination receiving retransmissions that are triggered either by the destination's immediate unsolicited negative acknowledgment or, if that or the corresponding retransmission is lost, by negative acknowledgments in the destination's status report in response to receiving a periodic poll message. Other aspects of SSCOP that make it suitable for high-speed optical networks include its 32bit aligned trailer-oriented protocol fields, and separation of control and information flow.

The size of ATM cells (53 bytes, including ATM overheads) is much smaller than that of conventional transport protocol segments [e.g., 1 KB (kilobyte)]. Furthermore, ATM does not recover from cell loss within the network and, in its native form, is oblivious to transport layer framing. This could lead to poor throughput of transport layers over ATM, as ATM could drop an early cell from one transport layer segment, and then waste resources on transmitting latter cells from that segment, and then discard a cell from the next segment, causing multiple segments to be damaged and need retransmission. This led to "ATM" switches being designed to be aware of transport layer framing. Switches that used the partial packet discard scheme [28] would continue discarding cells from a segment that had one of its cells discarded, confining loss to segments that would be retransmitted anyway. The early packet discard scheme [29] took this further, not only assuming that the transport layer would retransmit the whole segment, but assuming that the transport layer (like TCP) interpreted loss to indicate congestion, so when switch buffers were filling (but not yet overflowing), this scheme would discard all cells belonging to certain segments, preventing any partial packet delivery and prompting TCP's congestion avoidance, improving the aggregate throughput of TCP over ATM.

4.3. Xpress Transport Protocol

The Xpress Transport Protocol (XTP) [8] was developed in conjunction with the Protocol Engine project, with the aim of being a high-speed transport protocol that was suitable for VLSI implementation. The high speed of optical networks often motivates hardware implementations of protocols, and the next section of this article considers this topic in more depth.

Like SSCOP, XTP provides a "fastnak" mode, in which the destination immediately sends a negative acknowledgment in response to receiving a segment with a sequence number that does not follow its predecessor. This mode of operation is well matched to optical networks that tend to preserve sequence. XTP also provides the user with a choice of whether to use go-back-*n* or selective retransmission. The flow control of XTP contains both rate-based and window-based components.

5. IMPLEMENTATION ISSUES

Previous sections have addressed end-to-end protocol issues, whereas this section addresses how the transport protocol should be implemented in end systems. Proper implementation is important to ensure that transport protocols can handle the high transmission rates of optical networks.

The high transmission rates of optical networks encourage the use of large data units (packets and segments) in transport, and other protocols. Increasing the data unit size reduces the frequency with which per-dataunit operations (such as classification for demultiplexing and state lookup) need to be performed, and only impacts transmission efficiency if the application's payload is smaller than the transmission data unit. Consequently, optical networks often use packets that may be ≥ 8 KB "jumbograms." It is also desirable to avoid interrupting the destination processor at high speed as each data unit arrives because of the overhead in context switching. The interrupt frequency can be reduced by using *interrupt coalescing* [30], in which the processor is not interrupted for every packet received, but rather is interrupted for every *n* packets received (or after a short timeout).

The processing of packet headers can also be expedited by the format of the header reflecting processing requirements. For example, fields should be aligned on processor word boundaries whenever possible so that the processor does not need to shift them before operating on them. The packet header should also ideally fit within a processor cache line, and be aligned in memory with cache lines, so as to expedite header processing. Integer values in header fields may be represented in either little- or big-endian format. Ideally, the representation should match the native representation of the processors used in the end systems, and end systems may negotiate this representation as part of connection establishment [18]. The common alternative is to use a representation that is statically defined by the transport protocol, and for processors to adapt though translation. Since the headers of consecutive segments often differ little (e.g., perhaps in only a sequence number and checksum), they can be rapidly prepared at the source by copying the header from a template and then adjusting the fields that differ for this segment. Similarly, the destination can rapidly check the header of each incoming segment by comparing it to the header of the preceding segment [31].

Memory technologies have not increased in bandwidth as fast as optical transmission technologies, so highspeed transport protocols need to be implemented with appropriate memory management in order to match the rates of optical networks. This means minimizing the number of times that payload information is shifted in memory, such as using "buffer cutthrough" [32], in which data units remain stationary in memory when they are passed between layers of a protocol stack, with layers exchanging pointers to these data units, and encapsulating (or decapsulating) them in situ in memory.

Caching, as used to match memory speeds to increasing processor speeds, is of limited use for network interfaces, since they exhibit little temporal locality of reference. However, data sent over a network do exhibit spatial locality, and this makes video RAM technology [33] attractive for interfacing end systems to optical networks. Video RAMs consist of a large array of dynamic memory cells, and a fast static memory that can be rapidly stored or loaded with a row of the dynamic memory cells, or sequentially stored or loaded with values from off the chip. The conventional application of video RAMs is for the static memory to serially feed a videodisplay raster scan, while a processor can concurrently modify values in the dynamic memory. In the networking application [e.g., 8], the static memory feeds (or is fed by) the highspeed network interface, and the application accesses the dynamic memory.

Proper partitioning of the implementation of a transport protocol between hardware and software, and between the user-space and kernel-space software components can help a transport protocol keep up with optical networks. Functions that "touch" each byte that is transmitted (e.g., checksum calculations and encryption) are particularly appropriate for hardware implementation because of the high processor overhead of implementing these functions in software. If these functions must be implemented in software (e.g., so that they are readily changed), then the memory overhead of accessing each byte can be reduced by integrating these functions with functions from other layers of the protocol stack that also touch each byte [34]. The Protocol Engine project [8] attempted to implement complete protocol stacks, including the Xpress Transport Protocol, in hardware.

In addition to the partitioning of an implementation between hardware and software, there is also the issue of partitioning the software components of an implementation between user-space and kernel-space parts of the memory system. A transport protocol implementation needs to access services that an operating system provides, such as buffer management, scheduling, input/output access. Kernel-based transport protocols may reduces the overhead in accessing these services, increasing performance, but they are difficult to develop and deploy. Careful implementation can produce highperformance user-space implementations of transport protocols [e.g., 35].

BIOGRAPHY

Tim Moors is a Senior Lecturer in the School of Electrical Engineering and Telecommunications at the University of New South Wales, in Sydney, Australia. He researches transport protocols for wireless and optical networks, wireless LAN MAC protocols that support bursty voicestreams, communication system modularity, and fundamental principles of networking. Previously, he was with Polytechnic University in Brooklyn, New York, and the Communications Division of the Australian Defence Science and Technology Organisation. He received his Ph.D. and B.Eng.(Hons.) degrees from universities in Western Australia (Curtin and UWA).

BIBLIOGRAPHY

- W. Doeringer et al., A survey of light-weight transport protocols for high-speed networks, *IEEE Trans. Commun.* 38(11): 2025-2039 (Nov. 1990) (provides a classic survey of transport protocols for high-speed networks).
- International Telecommunications Union-T, Framework of Optical Transport Network Recommendations, ITU-T, Recommendation G.871, Oct. 2000.

TRANSPORT PROTOCOLS FOR OPTICAL NETWORKS 2621

- L. Kleinrock, The latency/bandwidth tradeoff in gigabit networks, *IEEE Commun. Mag.* 30(4): 36-40 (April 1992).
- 4. A. Jalali, P. E. Fleischer, W. E. Stephens, and P. C. Huang, 622 Mbps SONET-like digital coding, multiplexing, and transmission of advanced television signals on single mode optical fiber, *Proc. ICC*, April 1990, pp. 1043–1048.
- N. F. Maxemchuk, Problems arising from deflection routing: Live-lock, lockout, congestion and message reassembly, Proc. NATO Advanced Research Workshop on Architecture and Performance Issues of High-Capacity Local and Metropolitan Area Networks, June 1990, pp. 209–333.
- H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, IETF, RFC 1889, Jan. 1996.
- L. Zhang, Why TCP timers don't work well, *Proc. SIGCOMM*, Aug. 1986, pp. 397–405.
- G. Chesson, XTP/PE overview, Proc. 13th Conf. Local Computer Networks, Oct. 1988, pp. 292–296.
- D. D. Clark, M. L. Lambert, and L. Zhang, NETBLT: A high throughput transport protocol, *Proc. SIGCOMM*, Aug. 1987, pp. 353-359.
- D. R. Cheriton and C. Williamson, VMTP as the transport layer for high-performance distributed systems, *IEEE Commun. Mag.* 27(6): 37-44 (June 1989).
- T. R. Henderson, Design principles and performance analysis of SSCOP: A new ATM Adaptation Layer protocol, *Comput. Commun. Rev.* 25(2): 47–59 (April 1995).
- V. Jacobson, Congestion avoidance and control, Proc. SIG-COMM '88, Aug. 1988, pp. 314–329.
- J. H. Saltzer, D. P. Reed, and D. D. Clark, End-to-end arguments in system design, *Proc. 2nd Int. Conf. Distributed Computing Systems*, April 1981, pp. 509-512.
- J. D. Touch, Performance analysis of MD5, Proc. SIGCOMM, Aug. 1995, pp. 77–86.
- R. Stewart and C. Metz, SCTP: A new transport protocol for TCP/IP, *IEEE Internet Comput.* (Nov. 2001).
- D. L. Tennenhouse, Layered multiplexing considered harmful, Proc. Protocols for High Speed Networks, Nov. 1989, pp. 143-148.
- A. T. Campbell, G. Coulson, F. Garcia, and D. Hutchison, Resource management in multimedia communications stacks, Proc. IEE Int. Conf. Telecommunications, April 1993.
- S. Boecking, Object-Oriented Network Protocols, Addison-Wesley, Reading, MA, 2000.
- S. Murphy and A. Shankar, Service specification and protocol construction for the transport layer, *Proc. SIGCOMM*, Aug. 1988, pp. 88–97.
- B. Stiller, PROCOM: A manager for an efficient transport system, Proc. IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems (HPCS '92), Feb. 1992, pp. 0_14-0_17.
- A. Baratz et al., SNA networks of small systems, *IEEE J. Select. Areas Commun.* 3(3): 416–426 (May 1985).
- 22. V. Cerf, Y. Dalal, and C. Sunshine, Specification of Internet Transmission Control Program, IETF, RFC 675, 1974.
- V. Jacobson, R. Braden, and D. Borman, TCP extensions for High Performance, IETF, RFC 1323, May 1992.
- 24. M. Allman, V. Paxson, and W. Stevens, *TCP Congestion Control*, IETF, RFC 2581, April 1999.

2622 TRELLIS-CODED MODULATION

- M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, TCP Selective Acknowledgement Options, IETF, RFC 2018, Oct. 1996.
- K. Ramakrishnan, S. Floyd, and D. Black, *The Addition of Explicit Congestion Notification (ECN) to IP*, IETF, RFC 3168, Sept. 2001.
- W.-C. Feng, D. D. Kandlur, D. Saha, and K. G. Shin, Understanding and improving TCP performance over networks with minimum rate guarantees, *IEEE / ACM Trans. Network* 7(2): 173–187 (April 1999).
- G. Armitage and K. Adams, Packet reassembly during cell loss, *IEEE Network* 7(5): 26–34 (Sept. 1993).
- 29. A. Romanow and S. Floyd, Dynamics of TCP traffic over ATM networks, *Proc. SIGCOMM '94*, 1994.
- S. Muir and J. Smith, AsyMOS—an asymmetric multiprocessor operating system, Proc. OPENARCH, April 1998, pp. 25-34.
- D. Clark, V. Jacobson, J. Romkey, and H. Salwen, An analysis of TCP processing overhead, *IEEE Commun. Mag.* 27(6): 23–29 (June 1989).
- C. Woodside and J. Montealegre, The effect of buffering strategies on protocol execution performance, *IEEE Trans. Commun.* 37(6): 545-554 (June 1989).
- J. Nicoud, Video RAMs: Structure and applications, Proc. IEEE Micro, Feb. 1988, pp. 8–27.
- D. Clark and D. Tennenhouse, Architectural considerations for a new generation of protocols, *Proc. SIGCOMM 90*, Sept. 1990, pp. 200-208.
- A. Edwards, G. Watson, J. Lumley, and C. Calamvokis, Userspace protocols deliver high performance to applications on a low-cost Gbps LAN, *Proc. SIGCOMM*, Sept. 1994, pp. 14–23.
- 36. R. W. Watson and S. A. Mamrak, Gaining efficiency in transport services by appropriate design and implementation choices, ACM Trans. Comput. Syst. 5(2): 97-120 (May 1987) (Focuses on implementation issues).
- 37. J. P. G. Sterbenz and J. D. Touch, High-Speed Networking: A Systematic Approach to High-Bandwidth Low-Latency Communication, Wiley, New York, 2001 (provides in-depth coverage of high-speed networks).

TRELLIS-CODED MODULATION

STEPHEN G. WILSON University of Virginia Charlottesville, Virginia

1. INTRODUCTION

Coding for noisy channels has a half-century history of theory and applications, but until the early 1980s the applications usually presumed a binary-to-binary encoding function whose output was transmitted with a binary, say, phase-shift-keyed (PSK), modulation format. The emphasis was on achieving coding gain over an uncoded system, and it was understood that bandwidth expansion was one price to pay, often a tolerable price as in deep-space missions. In this regime, where bandwidth efficiency is less than 1 bps/Hz, finding strong codes amounts to finding codes with optimal binary Hamming distance properties.

On the other hand, most contemporary systems are pressed to communicate increasing throughput in limited bandwidth. Information theorists knew that substantial coding gain was also possible in the "bandwidth-efficient regime", where one is interested in communicating multiple bits per second per hertz. It was not until the seminal work of G. Ungerboeck [1] that this potential came to fruition, and thereafter the technique quickly penetrated applications. The general term for this channel coding technique is *trellis-coded modulation* (TCM). This article will present a tutorial overview of the main ideas behind TCM, with several examples. References to literature are given throughout for deeper investigation. In particular, Biglieri et al. [2] incorporate a wealth of more detailed information.

Coded modulation refers to the intelligent integration of channel coding and modulation to produce efficient digital transmission in the bandwidth-efficient regime. The essential theme is to map sequences of information bits to sequences of modulator symbols, these symbols drawn from a constellation having high spectral efficiency, in such a way as to maximize a relevant performance criterion. For the classic problem of signaling over the Gaussian channel, the objective becomes to maximize the minimum *Euclidean* distance between valid modulated sequences. The mapping can be block-oriented, called *block-coded modulation.*¹ TCM, on the other hand, is best viewed as a stream mapping, associating input bit (or symbol) streams with sequences of signal points from the modulator constellation.

The ideas generally are traced to the late 1970s and the work of Ungerboeck, although, as with most important innovations, earlier roots are evident (see, e.g., the preface to Ref. 2). The 1982 paper by Ungerboeck [1], following a 1976 Information Theory Symposium talk, prompted a great deal of subsequent research in TCM, and the technique quickly became part of the coding culture, appearing in several important voiceband modem and satellite communication standards and, more recently, wireless standards. It is relatively easy to obtain coding gains of several decibels relative to uncoded signaling for modest encoding/decoding complexity, without bandwidth expansion. For example, one of the designs discussed below sends 2 bits per modulator interval using 8-PSK modulation, with a 4-state encoder. The design achieves 3-dB asymptotic energy savings over uncoded QPSK, yet occupies the same bandwidth as the latter.

The basic themes common to TCM are rather simple. To achieve bandwidth efficiency in the first place, we need to communicate multiple bits per signal space dimension, since it can be shown that bandwidth occupancy is related to the number of complex signal space dimensions per unit time. To send k message bits per modulation interval, we adopt a constellation C containing more points than

¹ Block-coded modulation has never attained the foothold enjoyed by TCM in this application, probably because TCM represents a cleaner solution, and maximum-likelihood decoding is readily obtainable in TCM.

 2^k . (Typically, as seen shortly, the set size will be twice as large, viz., 2^{k+1} .) This constellation expansion provides the *redundancy* important to coded modulation, as opposed to simply sending extra points per unit time from a constellation of size 2^k . We divide the constellation into regular disjoint subsets, called cosets, such that the minimum Euclidean distance between members of these cosets is maximized (and greater than the original minimum distance of the constellation). k input bits are presented per unit time, and of these, $\tilde{k} \leq k$ bits are input to a finite-state encoder, having S states. This encoder produces a label sequence that picks a subset (or coset). The remaining k - k bits select the specific constellation point in the chosen subset. Signal points are produced at the same rate as input vectors are presented. This selection process has *memory* induced by the finite-state encoder, representing the other crucial aspect of coded communication. Figure 1 illustrates this generic encoding operation, sometimes called *coset coding* [3]. One might view this process as that of a time-varying modulation process, whereby the subset selector defines a signal set at each time interval from which the actual transmitted signal is to be selected. Of course, there are important details defining any particular design; in particular it is important to choose the proper value of k as well as the proper finite-state encoder.

In the next section, relevant information theory is summarized that points to the potential of TCM as well as proper design choices. Section 3 provides basic design principles with examples for simple TCM codes. Following that, more detailed information on performance evaluation and tables of best codes are provided. The article closes with a brief discussion of related topics, including design of codes for rotational invariance, multidimensional transmission, and fading channels.

2. RELEVANT INFORMATION THEORY

Before delving into the specifics of TCM, it is helpful to examine the potential efficiency gain that exists, according to information theory, for bandwidth-efficient communication. In addition, this study will provide insight into the appropriate choice of constellation size. This will be addressed for the important case of two-dimensional (in-phase/quadrature) modulation.

Consider the two-dimensional Gaussian noise channel shown in Fig. 2, where it is assumed that a signal vector $\mathbf{x} = (x_1, x_2)$ is presented at each channel use. This pair of



Figure 1. Depiction of generic TCM system.



Figure 2. Two-dimensional gaussian channel model.

real numbers corresponds also to a single complex input. The input is constrained only by its expected energy (i.e., $\mathcal{E}[x_1^2 + x_2^2] \leq E_s$), so that E_s represents the average symbol energy at the demodulator input. The noisy channel adds two-dimensional zero-mean Gaussian noise, with independent noise components in each coordinate. In keeping with standard notation, the variance of each component of the additive noise will be denoted $N_0/2$, where $N_0/2$ represents the power spectral density in watts per hertz of the physical white-noise process in the receiver.²

The channel capacity of such a channel is defined as the maximum of the mutual information between input and output vectors, the maximum taken over all probability assignments on input symbols satisfying the energy constraint. It is a standard result of information theory [4] that the capacity is attained when the input variables are independent Gaussian, with each subchannel allocated half the available energy. Moreover, the subchannel capacity is

$$C' = \frac{1}{2} \log_2 \left(\frac{1 + E'}{N_0/2} \right) \text{bits per channel use} \tag{1}$$

where E' is the energy available to each subchannel. Since the capacity of parallel independent channels equals the sum of the respective subchannel capacities and $E_s = 2E'$, we have

$$C = \log_2\left(\frac{1+E_s}{N_0}\right)$$
 bits per channel use (2)

Figure 3 plots this relation versus E_s/N_0 , scaled in dB (see the leftmost curve). Note that the plot, consistent with (2), is linear at high SNR, and every 3 dB increase in SNR buys an additional bit of channel capacity. On this same plot, we mark with circles the values of E_s/N_0

² This diagram encapsulates the actual physical process of waveform modulation, physical channel effects, and demodulation of waveforms to real numbers.



Figure 3. Capacity of two-dimensional signaling, AWGN channel.

needed to achieve bit error probability 10^{-5} when using uncoded 4-PSK, 8-PSK, 16-QAM, and 32-QAM (cross) constellations, techniques that are efficient designs for sending 1, 2, 3, or 4 bits per symbol, respectively. (See any text on digital communication, e.g., Ref. 5 or 6.) Noting that the capacity constraint implies the theoretical minimum E_s/N_0 capable of sending k bits per symbol, but that this limit is in principle closely approachable, we conclude that roughly 8–10 dB of potential efficiency improvement exists all across the high-rate regime, indicating significant potential exists for bandwidthefficient signaling, as well as in the classic coded binary signaling realm. This result was evident prior to the advent of TCM, but it is somewhat surprising that it took researchers so long to tap this potential.

To translate this into implications for bandwidthefficient signaling and for bottom-line energy efficiency limits, we argue that information rate R bits per channel use can be communicated reliably, provided R < C. When communicating R bits per symbol, we have $E_s = RE_b$, where E_b is the energy associated with an information bit, and obtain the relation

$$R \le \log_2\left(\frac{1+RE_b}{N_0}\right) \tag{3}$$

which thus requires that

$$\frac{E_b}{N_0} > \frac{2^R - 1}{R} \tag{4}$$

(This critical value is necessary for reliable communication, but is approachable by sufficiently sophisticated coding methods.) For example, to send R = 1 bit per 2D (two-dimensional) symbol, no matter what the constellation and no matter how complicated the coding process, it is impossible to achieve *reliable* communication if the bit SNR, E_b/N_0 , is less than 1, or 0 dB. More pertinent to our interest, if we wish to send R = 4 bits per interval, we must provide at least $E_b/N_0 > \frac{15}{4}$, or 5.7 dB. In terms of corresponding spectral efficiency, we can (optimistically) estimate that the required bandwidth needed to send R_s modulator symbols per second without intersymbol interference is $B = R_s$, approachable with Nyquist pulseshaping in 2D modulation. Thus, since $R_b = 4R_s$, we can communicate with (nearly) 4 bps/Hz spectral efficiency provided E_b/N_0 exceeds 5.7 dB.

In practice, code sequences are not fashioned from independent Gaussian random variables, although this result does illuminate the design of efficient TCM systems. Rather, the inputs to the channel are chosen from some finite, regular arrangement of M points (the constellation). Letting $P(\mathbf{x}_i)$ represent the probability of sending constellation point \mathbf{x}_i into the channel of Fig. 2, we can write that the mutual information between input and output is [4]

$$I(\mathbf{X}; \mathbf{Y}) = \sum_{i=0}^{M-1} P(\mathbf{x}_i) \int f(\mathbf{y} | \mathbf{x}_i) \log_2 \\ \times \left[\frac{f(\mathbf{y} | \mathbf{x})}{f(\mathbf{y})} \right] d\mathbf{y} \text{ bits/channel use}$$
(5)

This may be evaluated numerically by noting that the conditional PDFs are Gaussian 2D forms as described above. Normally, the input symbols are assumed to be equiprobable, which gives the "symmetric" capacity, although one should realize that choosing high-energy constellation points with smaller probability than inner, low-energy points, is a slightly superior choice mimicking the Gaussian distribution above. So-called shaping can extract some of this gain, perhaps amounting to 1 dB [7]. Nonetheless, the symmetric capacity for QPSK, 8-PSK, 16-QAM (quadrature amplitude modulation), and 32-QAM (cross-constellation) are shown in Fig. 3 as well. Notice that the capacity curves saturate, as expected, at $\log_2 M$, where M is the constellation size, since even without additive noise, $\log_2 M$, bits is the maximum error-free information per symbol.

An important conclusion from this analysis, drawn by Ungerboeck, is that to reliably communicate k bits per modulator symbol, constellations of size 2^{k+1} can achieve virtually the same energy efficiency as a hypothetical Gaussian code without constellation constraints would achieve. In specific terms, notice that the 8-PSK curve in Fig. 3 closely follows the Gaussian capacity curve up through capacity of 2 bits per symbol. Thus, to send k = 2 bits/symbol, 8-PSK represents a sensible choice. Whereas 16-QAM offers a greater selection of signals to build code sequences, the theoretical potential, expressed by channel capacity limits, is only incrementally better. To send k = 7 bits per interval, achieving even greater spectral efficiency, a sensible choice for modulation would be 256-QAM. (Note, however, that not any constellation will suffice; it is important that the constellation be an efficient packing of points into signal space.)

This "constellation doubling" is certainly convenient, for it represents the need for the encoder to produce one extra bit beyond the k input bits. The recommendation holds across the range of spectral efficiency, and is appropriate for constellations of any dimension, including 1D and 4D cases, for example.

3. TCM DESIGN PRINCIPLES

The discussion here will concentrate on 2D TCM, the case of primary practical interest. (A brief treatment of higherdimensional TCM is given at the end.) The underlying objective is to communicate k bits/modulator interval, thereby achieving a spectral efficiency approaching kbps/Hz. The ideas extend readily to higher dimensions though, and more will be said about the benefits of this later. Also, 1D, that is, pulse amplitude modulation (PAM), designs evolve easily from this basic formulation.

It is first helpful to understand what not to do in design. We should avoid the separation of the problem into design of a "good" binary encoder and the design of a "good" modulator for sending binary code symbols. In the case of k = 2, this strategy might locate tables of optimal Hamming distance codes, appropriate for binary modulation, with 2 input bits and 3 output bits per time step. Then we could adopt a Gray-labeled 8-PSK modulator for transmission of the 3 code bits. While such a design achieves the objective of sending 2 bits per modulator symbol, this decoupled approach cannot in general attain the same energy efficiency that a more integrated approach follows. (An exception is the case of binary codes mapped onto Gray-coded QPSK; there the squared Euclidean distance between signal points is proportional to Hamming distance between their bit labels, so optimal binary codes produce optimal TCM codes, and the bandwidth doubling normally attached to a rate- $\frac{1}{2}$ binary code, say, is avoided when mapping onto the larger QPSK constellation. In



some sense rate- $\frac{1}{2}\text{-coded}$ QPSK represents the earliest exemplar of TCM.)

3.1. Set Partitioning

Returning to the integrated design methodology, we adopt a constellation C that is a regular arrangement of $M = 2^m$ points in 2D signal space. We assume this constellation contains more than 2^k points, typically larger by a factor of 2 as discussed above. The constellation is first partitioned into disjoint subsets, A_i , whose union is the original constellation. For constellations we will discuss, this partition tower will involve successive steps of splittingby-2. Subsets are chosen so that the intraset minimum Euclidean distance (within subsets) is maximized. These subsets are often denoted as cosets, for one subset is merely a translation or rotation of another subset.

At the next level of the partition chain, each subset is further subdivided into smaller subsets, denoted \mathcal{B}_i , again increasing the intraset Euclidean distance. In principle the process continues until sets of size 1 are produced, although typically one does not need to proceed to this level.

The process is now illustrated for $8\mathchar`-PSK$ and $16\mathchar`-QAM$ constellations.

Example 1. Partitioning of 8-PSK. The 8-PSK constellation is comprised of 8 points equally spaced on a circle with radius $E_s^{1/2}$, so that E_s represents both the peak and average energy per modulator symbol. The partition shown in Fig. 4 first divides the constellation into two

Figure 4. Partition chain for 8-PSK.

QPSK sets, denoted $\mathcal{A}_0, \mathcal{A}_1$, one a rotation of the other. Notice that the intraset squared distance is $2E_s$ in normalized terms, whereas the original squared minimum distance is $d_0^2 = 0.586E_s$. Further splitting of each coset produces four antipodal (2-PSK) sets, within which the intraset squared distance is now $4E_s$. We denote these subsets by $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2$, and \mathcal{B}_3 . Finally, a third splitting produces singleton sets C_i .

Eventually, each constellation point will need to carry some binary label, 3 bits in this example. The partitioning process above provides a convenient means of doing so, and is called "mapping by set partitioning" by Ungerboeck [1]. In the partition chain, we have arbitrarily attached a 0 bit to a left branch and a 1 bit to a right branch at each stage. Reading the bits from bottom to top gives a 3-bit label to each point, as indicated in Fig. 4. Coincidentally, the bit labeling is the same as that of natural binary progression around the circle.

Example 2. Partitioning of 16-QAM. The partition chain for 16-QAM is depicted in Fig. 5, again showing a twofold splitting of sets at each stage. Here the squared distance growth is more regular than for M-PSK constellations specifically $4a^2$, $8a^2$, $16a^2$,..., where $E_s = 10a^2$ is the average energy per symbol for the constellation. This doubling behavior holds for arbitrary 2D constellations that are subsets of the integer lattice Z^2 . Note again the partitioning process supplies bit labels to each constellation point.

Other familiar constellations are easily partitioned in a similar manner, including 1D pulse amplitude modulation (PAM), M-ary phase shift keying, and large M-QAM sets. Identical methods pertain to partitioning of multidimensional lattice-based constellations as well, although the splitting factors are seldom twofold.

3.2. Trellis Construction

Given such a constellation partitioning, it remains to design a trellis code. The parameters of a trellis encoder are (1) the number of bits/symbol, k, as above, and (2) the number of encoder states, S, normally taken to be a power of 2, so $S = 2^{v}$. A trellis is merely a directed graph with S nodes (states) per timestep, and with 2^k edges joining each of these nodes to nodes at the next time stage. It is these edge label sequences that form the valid sequences of the trellis code. Already there exists some design choice, namely, how these 2^k edges emanating from each state connect with states at the next level. Equivalently, in Fig. 1, of the k input bits, how many will be used to influence the state of the encoder and how many remain to define a point within a subset. Suppose k = 2 and S = 8, for example. We might construct trellis graphs with each state branching to 4 distinct next states (k = 2), or, what is common in TCM designs, we could have two groups of "parallel" edges joining each state with two distinct states at the next level ($\tilde{k} = 1$). We call this "2 sets of 2" branching, versus "4 sets of 1" in the first case. The better policy is unknown at the outset, but the optimal choice becomes

 $d_0^2 = 4a^2$ \mathcal{A}_{f} \mathcal{A}_1 $d_1^2 = 8a^2$ \mathcal{B}_2 \mathcal{B}_1 \mathcal{B}_3 $d_2^2 = 16a^2$ $d_3^2 = 32a$ \mathcal{C}_4 \mathcal{C}_1 \mathcal{C}_3 \mathcal{C}_2 \mathcal{C}_{6} C_5 C_7

Figure 5. Partition chain for 16-QAM, $E_s = 10a^2$.

evident after study of all cases. In general, the trellis branching will become more diffuse when progressing to larger state complexity.

Given a choice of trellis topology we assign subsets of size $2^{k-\tilde{k}}$ to each state transition. These are commonly called *parallel edges* when $k - \tilde{k} \ge 1$. For small trellises it is not difficult to exhaustively try all possibilities, but for larger trellises there rapidly become many permutations of subset assignments, and their respective performances may differ.

Realizing that maximization of squared distance between distinct channel sequences is the objective, Ungerboeck proposed a "greedy" algorithm as follows:

- 1. Assign subsets to *diverging* edge sets in the trellis so that the interset minimum distance (between subsets) is maximized. This implies that two sequences that differ in their *state* sequences obtain maximal squared distance on the splitting stage.
- 2. Assign subsets to *merging* edge sets so that the interset minimum distance is maximized, for similar reason.

Further, all subsets are to be used equally often.

It may not be possible to achieve objectives 1 and 2 in small trellises, 2-state designs, for example. Also, this policy is only a heuristic that seems true of best codes found to date via computer search, but appears to have no provable optimality. Even within the class of "greedy" labelings there will remain many choices in general.

Example 3. 4-State TCM for 8-PSK, k = 2. The archetypal example of TCM designs is the 4-state code for 8-PSK, sending 2 bits per symbol. The trellis topology options are 2 sets of 2 ($\tilde{k} = 1$), or 4 sets of 1 ($\tilde{k} = 2$). Adopting the former, along with the greedy policy, and with reference to Fig. 4, we assign subsets of size 2, namely, \mathcal{B}_i , to the state transitions as shown in Fig. 6. One should note the symmetry present in the design—each constellation point is used exactly twice throughout the trellis, and each state has its exiting or entering arcs labeled with either $\mathcal{A}_0 = \mathcal{B}_0 \cup \mathcal{B}_2$ or $\mathcal{A}_1 = \mathcal{B}_1 \cup \mathcal{B}_3$. The greedy policy can also be seen — for example, sets \mathcal{B}_0 and \mathcal{B}_2 have maximal interset distance among the subset choices, and these are assigned to splits and merges at state 00.

$\begin{array}{c} D0 & D_0 \\ \hline B_2 \\ \hline D0 & B_1 \\ \hline B_3 \\ \hline B_2 \\ D1 & B_0 \\ \hline B_3 \\ \hline D1 & B_0 \\ \hline B_3 \\ \hline B_2 \\ \hline D1 & B_0 \\ \hline B_3 \\ \hline B_2 \\ \hline D1 & B_0 \\ \hline B_3 \\ \hline B_2 \\ \hline D1 & B_0 \\ \hline B_3 \\ \hline D1 & B_0 \\ \hline D1$

Figure 6. 4-state trellis for R = 2, 8-PSK.

It should be noted at this point that information bit sequences are not attached to trellis arcs, but the trellis only specifies valid sequences of modulator symbols. The focus is on maximizing the Euclidean distance between valid sequences, without concern as yet about the underlying message bit sequences.

3.3. Decoding of TCM

Before proceeding further on the design and performance aspects of TCM, we digress to describe optimal decoding, namely, maximum-likelihood sequence decoding.³ The TCM encoder produces a sequence c_i of 2D signal points for transmission over an additive Gaussian noise channel. Optimal decoding is easily understood in the context of Viterbi's algorithm for decoding the noisy output of any finite-state process [8,9]. The task is to find the most likely message sequence given the noisy observations, and the solution is the sequence whose sum of loglikelihoods for branch symbols is largest. For the Gaussian channel treated here, these branch metrics are merely the negative squared distances between the measurement and the constellation point being evaluated. Further simplification is possible if the constellation points have constant energy, as with *M*-PSK.

One may implement the decoder in an obvious manner by noting that each survivor state in the trellis such as Fig. 6 has 2^k branches entering it, and the survivor sequence to each state can then be computed by forming the cumulative metric for each of these 2^k sequences, and retaining the largest metric, as well as the route of the best path. In this view, decoding is little different from standard Viterbi processing. An alternate approach that exploits the structure of TCM first views the problem as finding the best-metric choice within each coset (the parallel edges of the graph), then having an add-compare-select routine evaluate the contending coset winners entering a given state. These coset winners can be found once and reused as needed for processing the remaining states at the same time index. Otherwise, aspects of the decoder are identical with standard Viterbi decoding. Issues of metric quantization and range, as well as survivor memory depth are relevant engineering issues, but will not be discussed here.

3.4. Design Assessment

We speak of error events as occurring when the decoder opts for some sequence other than that transmitted. These events generally correspond to short detours from the correct path, during which a small number of message bit errors occur. In contrast with conventional convolutional codes, 1-step error events often exist in TCM decoding. The fundamental parameter of interest in design of TCM codes is the *Euclidean free distance*, d_f , defined as the minimum, over all sequence pairs, of the Euclidean signal space distance between two valid code sequences, without regard to detour length. This is an extension of the notion

 $^{^{3}}$ This does not exactly minimize the decoded bit error probability, however.

of free distance for convolutional codes, where Hamming distance is the normal distance measure.

The possible error event lengths in the trellis of Fig. 6 are 1-step, along with 3-step, 4-step, and so on. Two-step error events are not possible in this case. For the present, assume that the transmitted sequence follows the top route in the trellis, corresponding to the "all-zeros path". A 1-step error event is the event that the decoder chooses the antipodal mate of the transmitted signal in a one-symbol measurement, and the squared distance for this pair is $4E_s$. (Recall that the radius of the constellation is $E_s^{1/2}$.) This is also the intraset squared distance for \mathcal{B}_0 .

There are four 3-step error events of the form 2-1-2, 2-1-6, 6-1-2, or 6-1-6, where the numbers represent subscripts of C_i in Fig. 4. Each sequence has squared distance from the 0-0-0 sequence of $2E_s + 0.586E_s + 2E_s = 4.586E_s$. It is straightforward, though tedious, to demonstrate that all 4-step, and longer, error events have distance of this code is $4E_s$, and we say 1-step error events dominate, since at high SNR, this kind of error event is more likely than any specific 3-step or longer error event. The probability of confusing two signals in Gaussian noise is given by

$$P_2 = Q \left[\left(\frac{d_E^2}{2N_0} \right)^{1/2} \right] \tag{6}$$

where $Q(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^{2}/2} dz$ is the Gaussian tail integral function. Given that each transmitted sequence has only one such nearest neighbor, and on each such error event, only 1 of the 2 information bits is in error, we predict that the asymptotic (high SNR) performance of this code is

$$P_b \approx \frac{1}{2} Q \left[\left(\frac{4E_b}{N_0} \right)^{1/2} \right] \tag{7}$$

We have also used the fact that the energy per symbol is $E_s = 2E_b$ (not $3E_b$). In comparison, the bit error probability of uncoded Gray-labeled 4-PSK is

$$P_{b_{\rm QPSK}} = Q \left[\left(\frac{2E_b}{N_0} \right)^{1/2} \right] \tag{8}$$

The *asymptotic coding gain* (ACG) is defined as the ratio of the arguments of the high-SNR error expressions for the coded and uncoded cases, and ignores multiplier constants. Here the ACG is 2, or 3.01 dB. The spectral occupancy remains unchanged, however, since we are sending one complex symbol for every 2 information bits in either case.

It is interesting to note that the error probability attached to the various message bits is in general not equal. In this example, the message bit defining which of the two parallel transitions is taken from a state to the next state is slightly more error-prone for large SNR. This is because the 1-step error event defines the free Euclidean distance for this code. We thus may find an unequal error protection property, which is sometimes taken as an opportunity if some message bits are deemed more important.

The encoder can be realized in more than one manner, in particular as a feedforward finite-state machine, or as one having feedback.⁵ Figure 7 illustrates two "equivalent" realizations of the 4-state encoder for 8-PSK. These encoders produce the same set of coded sequences, although the association between inputs and outputs differs. The 3-step error events attached to the first realization are produced by the lower input bit sequence 0001000..., while the same coded sequence is produced by the input 00010100... in the second encoder. Corresponding to this, the decoded *bit* error probability will also differ slightly. As with binary convolutional codes [10], we can always realize a TCM encoder as a systematic form with feedback, and this will be the emphasis from here on. Searching over this class of codes is convenient because in some sense this gives a minimal search space, and the encoders are automatically noncatastrophic as well, [1,2,6].

The alternative trellis topology, 4-sets-of-1 branching, cannot be made this efficient. Although 1-step error events are no longer possible, 2-step error events are present with squared distance inferior to the previous design, no matter the subset labeling on the trellis transitions. This latter configuration is, however, the topology associated with choice of an optimal 4-state binary encoder for maximizing Hamming distance—where parallel edges are seldom found.

When moving to an 8-state encoder, note that maintaining 2-sets-of-2 topology retains the 1-step error event, and thus the same free distance. (It is true that the longer error events become less problematic.) To increase the free distance, 1-step error events must be eliminated by switching to 4-sets-of-1 topology. The resulting optimal trellis is shown in Fig. 8, which, the reader may notice,



Figure 7. Encoders for 4-state TCM, R = 2, 8-PSK: (a) feedforward realization; (b) systematic, feedback realization.

 5 One should also realize that "different" realizations exist when the constellation labeling is changed.

⁴ A computer program can compute the minimum distance in more complicated situations, and would be used to evaluate codes in a code search.



Figure 8. 8-state trellis, R = 2, 8-PSK, dominant error events shown relative to upper path.

subscribes still to the greedy policy. The two dominant (3-step and 4-step) error events are shown, both with $d_f^2 = 4.586E_s$, and the asymptotic coding gain over QPSK is now 3.6 dB. Two-step events, although now present, have greater distance.

Example 4. 4-State Code for 16-QAM, k = 3. Another example that is feasible to design by hand is the 4-state code for 16-QAM. Each state now has $2^3 = 8$ entering and exiting branches, and the topology options are 2 sets of 4, or four sets of two. The former turns out to be best, again after studying the alternatives, and its trellis labeling is the same as that of Fig. 6, but uses the subset notation of Fig. 5. Since parallel edge sets are now size four, there are three 1-step error events relative to any transmitted sequence, but only two have the smallest squared distance, namely, the intraset minimum distance of \mathcal{B}_i in Fig. 5, $16a^2 = 1.6E_s = 4.8E_b$. Once again, there are no 2-step error events, and 3-step and longer error events have larger squared distance.

Following the methodology of the 8-PSK example, we can determine that for large SNR

$$P_b \approx N_b Q \left[\left(\frac{2.4E_b}{N_0} \right)^{1/2} \right] \tag{9}$$

where $N_b = \frac{1}{3}$ since message labels can be assigned so that at most one of the 3 input bits is incorrect on these two nearest sequences. In general this multiplier constant represents the average frequency of bit errors incurred among all the dominant distance error events, normalized by k, and depends on the code structure as well as the actual encoder realization.

This can be compared with the performance of uncoded 8-PSK, which is bounded by

$$P_b \le \frac{2}{3} Q \left[\left(\frac{0.88E_b}{N_0} \right)^{1/2} \right] \tag{10}$$

The asymptotic coding gain over uncoded 8-PSK (Graylabeled) is thus $10 \log_{10}(2.4/0.88) = 4.3$ dB. Again, this comparison is at equal spectral efficiency (3 bits/symbol); a comparison with uncoded 16-QAM would give slightly larger coding gain.

4. PERFORMANCE

Performance analysis for TCM resorts to upper, and perhaps lower, bounding of events known as the *node error* (or *first error*) *probability* and the *bit error probability*. The latter is generally of most interest, and the former is a useful step along the way. As with analysis of other coding techniques, we normally apply the union bound to obtain an upper bound, and this bound becomes tight at high SNR.

Suppose \mathbf{c}_i is an arbitrary transmitted sequence in a long trellis, and let \mathbf{c}_j be any other valid sequence in the trellis. Define \mathcal{I}_i to be the incorrect subset, namely, the set of valid trellis paths that split from \mathbf{c}_i at specific time k, and remerge at some later time. Then we have that the first-error-event probability is union-bounded by

$$P_{e} \leq \sum_{\mathbf{c}_{i}} P[\mathbf{c}_{i}] \sum_{\mathbf{c}_{j} \in \mathcal{I}_{i}} P[\Lambda(\mathbf{c}_{j}) > \Lambda(\mathbf{c}_{i})]$$
(11)

where $\Lambda(\mathbf{c}_i)$ is the total path metric for the sequence \mathbf{c}_i . Note that the bound is a sum of "2-codeword" error probabilities. These 2-codeword probabilities can be written for the AWGN channel as

$$P[\mathbf{c}_i \to \mathbf{c}_j] = Q\left[\left(\frac{d_E^2(\mathbf{c}_i, \mathbf{c}_j)}{2N_0}\right)^{1/2}\right]$$
(12)

where $d_E(\mathbf{c}_i, \mathbf{c}_j)$ is the Euclidean distance between the two sequences.

One important distinction for trellis codes is that the usual invariance to transmitted sequence may disappear; that is, the inner sum in (11) depends on the reference sequence \mathbf{c}_i , because either the sets of 2-codeword distances in the incorrect subsets vary, or more typically, the multiplicity at each distance varies with reference sequence. For linear binary codes mapped onto 2-PSK or 4-PSK, an invariance property holds that the allzeros sequence can be taken as the reference path, without loss of generality. A simple counterexample for TCM is provided by coded 16-QAM; transmitted sequences involving inner constellation points have a larger number of nearest-neighbor sequences than do transmitted sequences involving corner points, even though each has the same minimum distance to error sequences. Further details are not included here, but this issue is discussed under the topic of uniformity of the code, and conditions can be found for varying degrees of uniformity [2,11]. It turns out that the 8-PSK codes presented here are strongly uniform; that is, any transmitted sequence can be taken as a reference sequence. Other TCM schemes exhibit a weaker kind of uniformity in which every transmitted sequence has the same nearest-neighbor distance.

Given the general lack of a strong invariance property, the traditional transfer function bounding approach used to calculate (11) for convolutional codes must be generalized to average over sequence pairs. Reference 12 provides a graph-based means of doing this averaging, which provides numerical upper bounds on error event probability, and by differentiating the transfer function expression, bounds on decoded bit error probability (see also Ref. 2). These expressions are series expressions involving increasing effective distance, and as SNR increases, the leading (free distance) term in the expansion dominates.

The resulting expressions will be of the form

$$P_e \approx N_e Q \left[\left(\frac{d_E^2}{2N_0} \right)^{1/2} \right] \tag{13}$$

for error event probability and

$$P_b \approx N_b Q \left[\left(\frac{d_E^2}{2N_0} \right)^{1/2} \right] \tag{14}$$

for decoded bit error probability. The multiplier N_b can be interpreted as the average multiplicity of information bit errors over *all* error events whose distance equals the free distance, divided by the number of information bits released per trellis level, k. These expressions are the aforementioned dominant terms in the series expansions for error probability.

The multipliers either emerge from the transfer function bounding, retaining the dominant term, or, for simple codes, they may be found by counting. The k = 2 coded 8-PSK code with 4 states has $N_e = 1$ and $N_b = \frac{1}{2}$ as discussed earlier, but the multipliers can be significantly larger, especially for codes with many states.

One should be cautious in use of (13) or (14), for they represent neither a strict upper or lower bound. Instead, these are asymptotically correct expressions, tightening as SNR increases. Free Euclidean distance is not the entire story at high SNR, but the multiplier factor is also relevant, although a more second-order effect. It can be noted that in the vicinity of $P_b = 10^{-5}$, every factor of 2 increase in the multiplier coefficient costs roughly 0.2 dB in effective SNR for typical TCM codes.

5. TABLES OF CODES

In distinction with algebraic block code constructions, good TCM codes are produced by computer search that optimizes free Euclidean distance. In this section, we list properties of optimal codes taken from Ref. 13. Code information is listed for state complexities ranging from 4 to 64 states, deemed to be the range of most practical interest.

Table 1 summarizes data for the best 1D TCM codes for *M*-level PAM, where $M = 2^{k+1}$. Data are provided about (1) the systematic feedback encoder connection polynomials, using the notation of Fig. 9 and rightjustified octal form, where 13 corresponds to 001011; (2) the squared free distance normalized by $4a^2$, where

Table 1. Encoder Summary for Best PAM TCM Designs

					ACG(
States	${ ilde k}$	\mathbf{h}^1	\mathbf{h}^0	$d_f^2/4a^2$	k = 1, (1)	k = 2, (2)	N_e
4	1	2	5	9.0	2.6	3.3	4
8	1	04	13	10.0	3.0	3.8	4
16	1	04	23	11.0	3.4	4.2	8
32	1	10	45	13.0	4.2	4.9	12
64	1	024	103	14.0	4.5	5.2	36

Key: (1) gain relative to 2-PAM; (2) gain relative to 4-PAM. Source: Ref. 13.



Figure 9. General systematic trellis encoder with feedback.

2a is the PAM signal spacing along the real line; (3) the asymptotic coding gain over uncoded 2^k -ary PAM; and (4) the multiplier N_e in the first term of the union-bound expression for error event probability, as k becomes large. (In this case the same encoder operates on one input bit, regardless of k, at a given S.) The subset labels are 2 bits, meaning that the constellation is divided into four subsets. The remaining $k - \tilde{k}$ bits define a point from the selected subset. The first entry is the encoder used in the digital TV terrestrial broadcast standard in North America (see applications section below.)

Tables 2 and 3 provide similar information for 8-PSK and 16-PSK codes. Whereas 8-PSK is a good packing of 8 points in two dimensions, 16-PSK is not so attractive in an average energy sense, compared to 16-QAM. If peak energy (or amplitude) is of more concern, then 16-PSK improves [14]. The code described in Example 3 (above) is the 4-state code listed in Table 3, and it can be readily checked that the systematic form encoder of Fig. 7b concurs with the connection polynomials listed.

Table 2. Encoder Summary for Best 8-PSK TCM Designs, k = 2

States	\tilde{k}	\mathbf{h}^2	\mathbf{h}^1	\mathbf{h}^0	d_f^2/E_s	ACG (dB)	N_e
4	1	_	2	5	4.00	3.0	1
8	2	04	02	11	4.59	3.6	2
16	2	16	04	23	5.17	4.1	4
32	2	34	16	45	5.76	4.6	4
64	2	066	030	103	6.34	5.0	pprox 5.3

Key: (1) gain relative to 2-PAM; (2) gain relative to 4-PAM. Source: Ref. 13.

Table 3. Encoder Summary for Best 16-PSK TCM Designs, k = 3

States	\tilde{k}	\mathbf{h}^2	\mathbf{h}^1	\mathbf{h}^0	d_f^2/E_s	ACG (dB)	N_e
4	1	_	02	05	1.324	3.5	4
8	1		04	13	1.476	4.0	4
16	1	_	04	23	1.628	4.4	8
32	1	_	10	45	1.910	5.1	8
64	1	_	024	103	2.000	5.3	2

Key: (1) gain relative to 2-PAM; (2) gain relative to 4-PAM. Source: Ref. 13.

Table 4 compiles code and performance data for perhaps the most important case, 2D TCM using QAM constellations. As was the case in Table 1 for *M*-PAM, optimal codes with fixed state complexity share much in common as we increase k. For example in moving from R = 3 bits per symbol to R = 4, and so on, we can achieve this by simply growing the constellation by a factor of 2, adding one additional uncoded input bit and retaining the structure of the encoder for choosing cosets. This presents an attractive rate flexibility option.

Observe that for all 1D and 2D constellations, it is relatively easy to attain asymptotic coding gains of 3 dB to more than 5 dB with TCM, relative to an uncoded system with the same spectral efficiency. Notice also that for all these codes at most 2 of the k input bits influence the encoder state, regardless of state complexity. (This holds actually up through 256 states [13].) As k increases, the degree of parallelism in branching increases.

6. POWER SPECTRUM

If symbols are selected equiprobably and independently from a symmetric constellation (e.g., 16-QAM), the power spectrum of the transmitted signal does not contain spectral lines, and the continuum portion of the spectrum is given by the magnitude-squared of the Fourier transform of the modulator pulseshape [5,6]. A raisedcosine or root raised-cosine pulseshape is often selected to achieve band-limited transmission. The bandwidth of this spectrum scales according to the *symbol* rate, which is why high bandwidth efficiency accrues for large QAM signal constellations.

A standard approximation for the power spectrum of channel-coded signals is to adopt the spectrum of the uncoded modulation scheme, and frequency-scale according to the coded symbol rate, $R_{cs} = R_b/k$. This approximation models the coded symbol stream as an independent selection from the constellation, which, of course, is not valid. However, it turns out that for many coding techniques, the coded symbol stream exhibits pairwise independence; that is, the probability of successive pairs of symbols equals the product of the marginal probabilities. (This is not enough for strict independence.) This in turn implies that the coded stream is an uncorrelated one, and this is sufficient to yield a power spectrum identical to that of uncoded transmission, with care for scaling of the bandwidth according to number of information bits per symbol. In particular, codes proposed by Ungerboeck based on set partitioned labeling of the constellation and a linear convolutional encoder behave this way [15].

Example 5. Power Spectrum for Satellite Transmission. High-speed data transmission via satellite at 120 Mbps might utilize R = 2 coded 8-PSK transmission, so the symbol rate is 60 Msps (60 million symbols per second). Using root raised-cosine pulseshaping with rolloff factor 0.2, the total signal bandwidth occupies a range of 72 MHz, consistent with certain satellite transponder bandwidths.

7. APPLICATIONS

In this section, two contemporary applications of TCM are described. The actual transmission rates differ markedly, with one intended for dialup modems over the telephone network, while the other supports RF broadcast of digital TV. Nonetheless the motivation and operating principles are similar.

7.1. ATSC Television

The North American standard for terrestrial broadcasting of digital television goes under the name of the American Television Standards Committee (ATSC) standard [16]. There are numerous modes of operation, but one employs 4-state TCM encoding of 8-level PAM (a 1D TCM system). Coded symbols are passed through pulseshaping filters and vestigial sideband modulation is employed to keep the signal bandwidth within a 6-MHz allocation, even though the PAM symbol rate is 10.7 Msps.

The 4-state encoder is shown in Fig. 10. Since only 1 of the 2 input bits influences the encoder state, the $2^2 = 4$ branches leaving each state are organized into 2 sets of

Table 4. Encoder Summary for Best TCM Designs for QAM

						ACG (dB)				
States	${ ilde k}$	\mathbf{h}^2	\mathbf{h}^1	\mathbf{h}^0	$d_f^2/4a^2$	(k = 3),(1)	(k = 4),(2)	(k = 5),(3)	N_e	
4	1	_	02	05	4.0	4.4	3.0	2.8	4	
8	2	04	02	11	5.0	5.3	4.0	3.8	16	
16	2	16	04	23	6.0	6.1	4.8	4.6	56	
32	2	10	06	41	7.0	6.1	4.8	4.6	16	
64	2	064	016	101	8.0	6.8	5.4	5.2	56	

Key (1) gain relative to 8-PSK; (2) gain relative to 16-QAM; (3) gain relative to 32-QAM. Source: Ref. 13.



Figure 10. Trellis encoder for ATSC digital TV standard.

2, as for the trellis of Example 3 above. However, here the 1-step error event is not dominant due to the large intraset distance between points in sets of size 2. It can be shown that a 3-step error event dominates the distance, with free Euclidean distance $d_f^2 = \frac{36}{21}E_s$, leading to a 3.3-dB asymptotic coding gain over four-level PAM, (see Table 1).

Coherent detection is performed in the receiver with the aid of a pilot carrier to achieve phase lock without phase ambiguity. The TCM system is actually concatenated with an outer Reed–Solomon code to further improve performance on noisy channels. Channel equalization for multipath effects is also important.

7.2. V.32 Modem

One of the earliest applications of TCM appeared in the V.32 modem standard for 9600-bps transmission over the dialup telephone network. Prior to this time, all modem standards utilized uncoded transmission of modulator symbols. Use of 16-QAM, for example, can achieve 9600 bps with a symbol rate of 2400 sps, consistent with the voiceband channel bandwidth. Of course this data speed is now regarded as slow, but similar ideas have pushed speeds up by a factor of 3 over the same media [17].

To achieve the same throughput without increase in bandwidth, the V.32 standard employs a TCM design due to Wei [18] that maps onto 32-QAM with an 8-state encoder shown in Fig. 11 along with the constellation labeling of Fig. 12. In this trellis, 2 of the 4 bits entering the encoder do not influence the state vector, and are called *uncoded bits*. The remaining 2 are differentially encoded to handle rotational ambiguity (see Section 8), and the differential encoder output influences the state sequence, and hence the sequence of cosets. Each state has 16 branches emanating, organized as 4 sets of 4. In this trellis there are 1-step, 2-step, 3-step, and longer error events. The dominant error event(s) are 3-step events, and the asymptotic coding gain, relative to uncoded 16-QAM, is 4 dB.

Notice that the encoder is nonlinear over the binary field, due to the AND gates, but this does not complicate the decoder relative to a linear encoder. It is also noteworthy that this particular code has the same asymptotic coding gain as the best non-rotationally invariant code with 8 states (Table4). Normally, this extra constraint implies a small distance penalty, however.

A simulation of the performance of the V.32 design was performed with 10^6 bits sent through the system at each SNR in 1-dB steps from 6 to 11 dB. Results for decoded bit error probability, after differential decoding, are shown in Fig. 13, along with a plot of the expression $50Q[(2E_b/N_0)^{1/2}]$. The argument of the Q-function is obtained from the fact that the coding gain over uncoded 16-QAM is 4.0 dB, and that uncoded 16-QAM is 4.0 dB inferior in distance to uncoded QPSK. Thus the asymptotic energy efficiency is equivalent to that of uncoded QPSK, vet the system sends 4, rather than 2, bits per 2D symbol. The factor 50 is an empirically determined value that seems to fit the data well, and is consistent with the multipliers found in Table 4, given that multiple bit errors may occur per error event, and that the differential decoder increases the final bit error probability.

8. FURTHER TOPICS

8.1. Rotational Invariance

To perform coherent (known carrier phase) detection, the receiver must estimate the carrier's phase/frequency from the noisy received signal. This is normally done by some sort of feedback tracking loop with decisiondirected operation to remove the influence of data symbols. However, given a symmetric constellation in one or two dimensions, this estimator will exhibit a phase ambiguity of $0/180^{\circ}$ or $0/90/180/270^{\circ}$, respectively. Essentially, without additional side information, the demodulator has no way of discriminating phase beyond an ambiguity implied by the symmetry order of the constellation.

To resolve this ambiguity without first making hard decisions on symbols, then performing differential decoding prior to trellis decoding, a *rotationally invariant*



Figure 11. Trellis encoder for V.32 modem standard [18].



Figure 13. Performance of V.32 modem standard.

design is often preferred. This is a TCM design for which (1) a rotated version of every valid code sequence is also in the code space and (2) all such rotated versions correspond to the same information sequence, when launched from a different initial state. Usually a somewhat weaker condition is imposed—namely, that all rotated versions of the same code sequence correspond to encoder input sequences that, when differentially decoded, correspond to the same information sequence. In the latter case, the encoder is preceded by a precoder on some set of input bits, and the output of the TCM decoder is processed by a differential decoder acting on these same bits, thereby resolving the ambiguity. The penalty for not knowing phase outright is a slight increase in bit error probability.

Conditions for rotational invariant design were studied in detail by Trott et al. [19]. A notable example of a RI design is the 8-state TCM code for 32-QAM modulation due to Wei [18]. The encoder is depicted in Fig. 11 as described above. Figure 12. Constellation labeling for V.32 modem standard [18].

8.2. Multidimensional TCM

The codes described thus far produce symbol streams of one- or two-dimensional constellation symbols. Slightly more efficient designs are available when we treat the constellation as a four-dimensional (or larger) object. The typical means of fashioning a four-dimensional constellation is to use two consecutive 2-D symbols. If C is 2D, then the set product $C \times C$ is four-dimensional.

This higher-dimensional constellation can be partitioned in a manner similar to that described earlier for 2D constellation, and then subsets are assigned to trellis arcs so that Euclidean distance is maximized. To keep the throughput fixed, however, one must realize that subsets are much larger in this case. For example, if a 2D TCM system is to send 3 bits per 2D symbol, say, using 16-QAM, then in a 4D TCM design, the encoder sends 6 bits per pair of 2D symbols. If the encoder has, say, 8 states, with arcs to 2 other states, then each arc must be labeled with a subset of size 32 (4D) points.

There are a few advantages offered by higherdimensional codes, although they should be understood as second-order improvements: (1) it is possible to achieve slightly larger coding gains when throughput and trellis complexity are fixed; (2) rotationally invariant codes are easier to obtain in higher dimensions; and (3) Finally, one may exploit constellation shaping to better advantage, whereby the multi-dimensional constellation is more spherical, and 2D cross sections of these constellations are smaller than the corresponding constellation for 2D coding. Eyuboglu et al. [17] provide an example of a modem standard that employs multidimensional TCM together with shaping. Additional discussion on multidimensional TCM designs can be found in Refs. 13 and 20.

8.3. TCM on Fading Channels

Some propagation channels, notably wireless channels between terminals among buildings or influenced by terrain effects, are beset with the phenomenon of fading. Essentially, the multiple paths for energy from transmitter to receiver may combine in a constructive or destructive manner, varying with time assuming some motion of the physical process. A common assumption, more accurate in the case of large numbers of scattering objects, is that the aggregate signal amplitude obeys a Rayleigh distribution, leading to the Rayleigh fading assumption. Normally, the fading is "slow"; that is, the amplitude is essentially fixed over many consecutive symbols. Traditional TCM coding does not fare well in such cases, for a single span of faded code symbols easily defeat the power of the decoder to combat noise.

If the fading effect can be made to appear essentially independent from symbol to symbol, on the other hand, then TCM can achieve *diversity* protection against fading, essentially mimicking a time diversity technique without the loss of throughput the latter implies. One way to achieve this independence is via interleaving, either in block or convolutional manner, to sufficient depth. (On slow-fading media, this may be impossible if constraints on delay are tight, as they are in two-way voice communication.) The encoder output is interleaved prior to modulation, and the demodulator output, together with a channel amplitude estimate, is deinterleaved prior to TCM decoding. The amplitude estimate is used in building the branch metric for the various trellis levels according to

$$\lambda = r_n a_n^* x_n^* \tag{15}$$

where a_n is the complex channel gain, assumed known, at time n, r_n is the complex channel measurement at time n, and x_n is the signal being evaluated. Recalling that the union bound on TCM performance involves a sum of 2-codeword probabilities, we are interested in minimizing the 2-codeword probability for all sequence pairs in the face of independent Rayleigh variables. It may be shown that the 2-codeword probability for confusing two sequences in independent Rayleigh fading, assuming perfect CSI (channel state information), is, at high SNR [21,22]

$$P(\mathbf{c}_1 \to \mathbf{c}_2) \le \frac{c}{(E_b/N_0)^D} \tag{16}$$

where *D* is the *Hamming* symbol distance between the two sequences. Here *c* is a proportionality factor dependent on the constellation and code, and E_b should be interpreted as the average received energy per bit.

This implies that the design criterion should (1) maximize the minimum Hamming distance between sequences (as opposed to Euclidean distance) and (2) within this class of codes, maximize the product of Euclidean distances between symbols on sequences on the minimum Hamming distance pairs, the so-called product distance [21]. This gives much different codes than the AWGN optimization provides. For example, a 4-state code for 8-PSK, to be used on the interleaved Rayleigh channel, should not have parallel transitions as in Fig. 6, but instead should have a 4-sets-of-1 trellis topology [21,22]. Here the minimum Hamming distance between sequences is 2, and the TCM system provides dual diversity.

Another approach to providing diversity is to use multiple symbols per trellis branch, amounting to

multidimensional signaling. Continuing the previous case, instead of sending 2 bits per single constellation symbol, we could agree to send 4 bits per pair of symbols, keeping the rate the same. This so-called *multiple trellis-coded modulation* can achieve higher diversity order in some cases, as well as achieve greater product distance [23].

Another approach is to employ bit interleaving, rather than symbol interleaving. Here the objective is to design the code for maximal binary Hamming distance, rather than symbol distance. Readers are referred to Ref. 24 for further information in this regard.

Acknowledgments

The author gratefully acknowledges the assistance of Griffin Myers in preparing this work and the helpful comments of William Ryan.

BIOGRAPHY

Stephen Wilson is currently Professor of Electrical Engineering at the University of Virginia, Charlottesville, Virginia. His research interests are in applications of information theory and coding to modern communication systems, specifically digital modulation and coding techniques for satellite channels and wireless systems; spread spectrum technology; wireless antenna arrays; transmission on time-dispersive channels; and software radio. Prior to joining the University of Virginia faculty, Dr. Wilson was a staff engineer for The Boeing Company, Seattle, Washington, engaged in system studies for deepspace communication, satellite air-traffic-control systems, and military spread spectrum modem development. Prof. Wilson is presently area editor for Coding Theory and Applications of the IEEE Transactions on Communications, and the author of the graduate-level text Digital Modulation and Coding. He also acts as consultant to several industrial organizations in the area of communication system design and analysis and digital signal processing.

BIBLIOGRAPHY

- G. Ungerboeck, Channel coding with amplitude/phase modulation, *IEEE Trans. Inform. Theory* IT-28: 55-67 (Jan. 1982).
- E. Biglieri, D. Divsalar, P. McLane, and M. K. Simon, Introduction to Trellis Coded Modulation with Applications, Macmillan, New York, 1991.
- G. D. Forney, Jr., Coset codes part 1: Introduction and geometrical classification, *IEEE Trans. Inform. Theory* IT-34: 1123-1151 (Sept. 1988).
- T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
- 5. J. Proakis, *Digital Communications*, McGraw-Hill, New York, 2001.
- S. Wilson, Digital Modulation and Coding, Prentice-Hall, New York, 1996.
- G. D. Forney, Jr., Trellis shaping, *IEEE Trans. Inform.* Theory IT-38: 281-300 (1992).
- A. J. Viterbi, Error bounds for convolutional codes and asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* IT-13: 260-269 (1967).

- G. D. Forney, Jr., The Viterbi algorithm, *IEEE Proc.* 61: 268–278 (1973).
- G. D. Forney, Jr., Convolutional codes I: Algebraic structure, IEEE Trans. Inform. Theory IT-16: 720–738 (1970).
- G. D. Forney, Jr., Geometrically uniform codes, *IEEE Trans. Inform. Theory* IT-37: 1241–1260 (1991).
- E. Zehavi and J. K. Wolf, On the performance evaluation of trellis codes, *IEEE Trans. Inform. Theory* **IT-32**: 196–202 (March 1987).
- G. Ungerboeck, Trellis coded modulation with redundant signal sets, parts 1 and 2, *IEEE Commun. Mag.* 25(2): 5–21 (Feb. 1987).
- S. G. Wilson, P. J. Schottler, H. A. Sleeper, and M. T. Lyons, Rate 3/4 trellis coded 16-PSK: Code design and performance evaluation, *IEEE Trans. Commun.* COM-32: 1308–1315 (Dec. 1984).
- E. Biglieri, Ungerboeck codes do not shape the power spectrum, *IEEE Trans. Inform. Theory* **IT-32**: 595–596 (July 1986).
- 16. J. Whitaker, DTV Handbook, McGraw-Hill, New York, 2001.
- M. V. Eyuboglu, G. D. Forney, Jr., P. Dong, and G. Long, Advanced modulation techniques for V. fast, *Eur. Trans. Telecommun.* 4: 243-256 (May 1993).
- L. F. Wei, Rotationally invariant convolutional channel coding with expanded signal space, part II: Nonlinear codes, *IEEE J. Select. Areas Commun.* SAC-2: 672-686 (Sept. 1984).
- M. D. Trott, S. Benedetto, R. Garello, and M. Mondin, Rotational invariance of trellis codes: Encoders and precoders, *IEEE Trans. Inform. Theory* IT-42: 751-765 (1996).
- S. S. Pietrobon et al., Trellis-coded multidimensional phase modulation, *IEEE Trans. Inform. Theory* **IT-36**: 63–89 (Jan. 1990).
- D. Divsalar and M. K. Simon, The design of trellis-coded MPSK for fading channels: Set partitioning for optimum code design, *IEEE Trans. Commun.* COM-36: 1004-1011 (Sept. 1988).
- 22. Y. S. Leung and S. G. Wilson, Trellis coding for fading channels, in *ICC Conf. Record* (Seattle), 1987.
- 23. D. Divsalar and M. K. Simon, Multiple trellis-coded modulation, *IEEE Trans. Commun.* COM-36: 410-419 (April 1988).
- G. Caire, G. Taricco, and E. Biglieri, Bit-interleaved coded modulation, *IEEE Trans. Inform. Theory* **IT-44**: 927–946 (1998).

TRELLIS CODING

CHRISTIAN SCHLEGEL University of Alberta Edmonton, Alberta, Canada

1. INTRODUCTION

The tremendous growth of high-speed logic circuits and very large-scale integration (VLSI) has ushered in the digital information age, where information is stored, processed, and moved in digital format. Among other advantages, digital signals possess an inherent robustness in noisy communications environments. If distorted, they can be restored, and, through signal processing techniques, errors that occur during transmission can be corrected. This process is called *error control coding*, and is accomplished by introducing dependencies among a large number of digital symbols. *Trellis coding* is a specific, widespread error control methodology.

Figure 1 shows the basic configuration of a point-topoint digital communications link. The digital data to be transmitted over this link typically consists of a string of binary symbols: ones and zeros. These symbols enter the encoder/modulator whose function it is to prepare them for transmission over a channel, which may be any one of a variety of physical channels. The encoder accepts the input digital data and introduces controlled redundancy for transmission over the channel. The modulator converts discrete symbols given to it by the encoder into waveforms that are suitable for transmission through the channel. On the receiver side, the demodulator reconverts the waveforms back into a discrete sequence of received symbols, and the decoder reproduces an estimate of the digital input data sequence, which is subsequently used by the data sink. The purpose of the error control functions is to maximize data reliability when transmitted over an unreliable channel.

An important auxillary function of the receiver is synchronization, that is, the process of acquiring carrier frequency and phase, and symbol timing in order for the receiver to be able to operate. Compared to data detection, synchronization is a relatively slow process, and therefore we usually find these two operations separated in receiver implementations.

Another important mechanism in many communication systems is *automatic repeat request* (ARQ). In ARQ the receiver additionally performs error detection, and, through a return channel, requests retransmission of data blocks that cannot be reconstructed with sufficient confidence. ARQ can usually improve the data transmission quality substantially, but a return channel, which is needed for ARQ, is not always available, or may be impractical. For a deep-space probe, for example, ARQ is infeasible since the return path takes too long (several hours!). Equally so, for speech encoded signals ARQ is usually infeasible since only a maximum speech delay of 200 ms is acceptable. In broadcast systems, ARQ is ruled out for obvious reasons.

Error control coding without ARQ is termed *forward error control* (FEC) coding. FEC is more difficult to perform than simple error detection and ARQ, but dispenses with the return channel. Often, FEC and ARQ are combined in hybrid error control systems.

2. IMPORTANCE OF FORWARD ERROR CONTROL (FEC) CODING IN A DIGITAL COMMUNICATIONS SYSTEM

The modern approach to error control coding in digital communications started with the groundbreaking work of Shannon [1], Hamming [2], and Golay [3]. While Shannon advanced a theory to explain the fundamental limits on the efficiency of communications systems, Hamming and Golay were the first to develop practical error control schemes. The new revolutionary paradigm born was



Figure 1. System diagram of a complete point-topoint communication system for digital data.

Figure 2. Popular 2D signal constellations used for digital radio systems.

one in which errors are no longer synonymous with data that are irretrievably lost, but by clever design, errors could be corrected, or avoided altogether. Although Shannon's theory promised that large improvements in the performance of communication systems could be achieved, practical improvements had to be excavated by laborious work. In the process, coding theory has evolved into a flourishing branch of applied mathematics [4].

The starting point to coding theory is Shannon's celebrated formula for the capacity of an ideal band-limited Gaussian channel, given by

$$C = W \log_2\left(1 + \frac{S}{N}\right)$$
 bps (bits per second) (1)

In this formula C is the channel capacity, which is the maximum rate of information, measured in bits per second, which can be transmitted through this channel; W is the bandwidth of the channel, and S/N is the signal-tonoise power ratio at the receiver. Shannon's main theorem, which accompanies Eq. (1), asserts that error probabilities as small as desired can be achieved as long as the transmission rate R through the channel (in bits per second) is smaller than the channel capacity C. This can be achieved by using an appropriate encoding and decoding operation. On the other hand, the converse of this theorem states that for rates R > C there is a significant error rate which cannot be reduced no matter what processing is invoked.

2.1. Bandwidth and Power

In order to be able to appreciate fully the concepts of trellis coding and trellis-coded modulation, some signal basics are needed. Nyquist showed in 1928 [25] that a channel of bandwidth W (in Hertz) is capable of supporting approximately 2W independent signal dimensions per second. If two carriers $[\sin(2\pi f_c) \text{ and } \cos(2\pi f_c)]$ are used in quadrature, as in double-sideband suppressed-carrier (DSBSC) amplitude modulation, we alternatively have W pairs of dimensions (or complex dimensions) per second, leading to the popular QAM (quadrature amplitude modulation) constellations, represented by points in two-dimensional (2D) space. Some popular constellations for digital communications are shown in Fig. 2.

The parameter that characterizes how efficiently a system uses its allotted bandwidth is the *spectral efficiency* η , defined as

$$\eta = \frac{\text{bit rate}}{\text{channel bandwidth } W} \qquad (\text{bps/Hz}) \qquad (2)$$

Using (1) and dividing by W, we obtain the maximum spectral efficiency for an additive white Gaussian noise (WGN) channel, the *Shannon limit*, as

$$\eta_{\max} = \log_2\left(1 + \frac{S}{N}\right)$$
 (bps/Hz) (3)

To calculate η , we must suitably define the channel bandwidth W. One commonly used definition is the 99%



bandwidth definition, where W is defined such that 99% of the transmitted signal power falls within the band of width W. This 99% bandwidth corresponds to an out-of-band power of -20 dB.

The average signal power S can be expressed as

$$S = \frac{kE_b}{T} = RE_b \tag{4}$$

where E_b is the energy per bit, k is the number of bits transmitted per symbol, and T is the duration of that symbol. The parameter R = k/T is the transmission rate of the system in bits per second. Rewriting the signalto-noise power ratio S/N, where $N = WN_0$, where total noise power equals the noise power spectral density N_0 multiplied by the width of the transmission band, we obtain

$$\eta_{\max} = \log_2\left(1 + \frac{RE_b}{WN_0}\right) = \log_2\left(1 + \eta\frac{E_b}{N_0}\right) \tag{5}$$

Since $R/W = \eta_{\text{max}}$ is the limiting spectral efficiency, we obtain a bound from (5) on the minimum bit energy required for reliable transmission, given by

$$\frac{E_b}{N_0} \ge \frac{2^{\eta_{\max}} - 1}{\eta_{\max}} \tag{6}$$

which is also called the Shannon bound.

In the limit as we allow the signal to occupy an infinite amount of bandwidth, that is, $\eta_{\text{max}} \rightarrow 0$, we obtain

$$\frac{E_b}{N_0} \ge \lim_{\eta_{\max} \to 0} \frac{2^{\eta_{\max}} - 1}{\eta_{\max}} = \ln(2) = -1.59 \text{ dB}$$
(7)

Figure 3. Bit error probability of quadrature phase shift keying (QPSK) and selected 8-PSK trellis-coded modulation (TCM) methods as a function of the normalized signal-to-noise ratio.

the minimum bit energy to noise power spectral density required for reliable transmission.

2.2. Communications System Performance with FEC

In order to compare different communications systems, a second parameter, expressing the power efficiency, has to be considered also. This parameter is the information bit error probability P_b . Figure 3 shows the error performance of QPSK, a popular modulation method for satellite channels that allows data transmission of rates up to 2 bps/Hz (bits per second per Hertz). The bit error probability of QPSK is shown as a function of the signalto-noise ratio S/N per dimension normalized per bit, henceforth called SNR. It is evident that an increased SNR provides a gradual decrease in error probability. This contrasts markedly with Shannon's theory, which promises zero(!) error probability at a spectral rate of 2 bps/Hz, if SNR > 1.5 (1.76 dB). The dashed line in the figure represents the Shannon bound adjusted to the bit error rate.

Also shown in Fig. 3 is the performance of several trellis-coded modulation (TCM) schemes using 8-ary phase-shift keying (8-PSK), and the improvement of coding becomes evident. The difference in SNR for an objective target bit error rate between a coded system and an uncoded system is termed the *coding gain*. It is important to point out here that TCM achieves these coding gains without requiring more bandwidth than the uncoded QPSK system. The figure also shows the performance of a more recently proposed method, Turbo trellis-coded modulation (TTCM). This extremely powerful coding scheme comes very close to the Shannon limit.

Figure 4. Measured bit error probability of QPSK and a 16-state 8-PSK (TCM) modem over a 64-kps satellite channel [10].

As discussed later, a trellis code is generated by a circuit with a finite number of internal states. The number of these states is a measure of its decoding complexity if optimal decoding, that is, *maximum-likelihood* (ML) decoding, is used. However, the two very large codes shown are not decoding via optimal decoding methods; they are sequentially decoded (see Schlegel's [5] discussion of Wang and Costello's paper [7]). TTCM is the most complex coding scheme, requiring not only two trellis decoders but also soft information decoding and storage of large blocks of data. FEC realizes the promise of Shannon's theory, which states that for a desired error rate of $P_b = 10^{-6}$ we can gain almost 9 dB in expended signal energy with respect to QPSK.

In Fig. 4 we compare the performance of a 16-state 8-PSK TCM code used in an experimental implementation of a single-channel-per-carrier (SCPC) modem operating at 64 kbps (1000 bits per second) [10] against QPSK and the theoretical performance established via simulations. As an interesting observation, the 8-PSK TCM modem comes much closer to its theoretical performance than the original QPSK modem, and a coding gain of 5 dB is achieved.

Figure 5 shows the performance of selected convolutional codes on an additive white Gaussian noise channel. Contrary to TCM, convolutional codes (with BPSK modulation) do not preserve bandwidth and the gains in power efficiency in Fig. 5 are partly obtained by a power bandwidth tradeoff; specifically, the rate $\frac{1}{2}$ convolutional codes plotted here require twice as much bandwidth as does uncoded transmission. This bandwidth expansion may not be an issue in deep-space communications and the application of error control to spread-spectrum systems. As a consequence, for the same complexity, convolutional codes achieve a higher coding gain than does TCM. Turbo coding, the most complex of the coding schemes, achieves



a performance within fractions of 1 dB of the Shannon limit.

The field of error control and error-correction coding somewhat naturally breaks into two disciplines, namely, block coding and trellis coding. While block coding, which is approached mostly as applied mathematics, has produced the bulk of publications in error control, trellis coding seems to be favored in most practical applications. One reason for this is the ease with which soft-decision decoding can be implemented for trellis codes. Soft decision is the operation when the demodulator no longer makes any (hard) decisions on the transmitted symbols, but passes the received signal values directly on to the decoder. The decoder, in turn, operates on reliability information obtained by comparing the received signals with the possible set of transmitted signals. This gives soft decision a 2-dB advantage. Also, trellis codes are better matched to high-noise channels, that is, their performance is less sensitive to SNR variations than the performance of block codes. In many applications the trellis decoders act as "SNR transformers"; that is, they lower the signalto-noise ratio from the input to the output. Such SNR transformers find application in Turbo decoders, coded channel equalizers, and coded multiple-access systems. The irony is, that in many cases, block codes can be decoded more successfully using methods developed originally for trellis codes than using their specialized decoding algorithms.

Figure 6 shows the power and bandwidth efficiencies of some popular uncoded quadrature constellations as well as that of a number of coded transmission schemes. The plot clearly demonstrates the advantages of coding. The trelliscoded modulation schemes used in practice, for example, achieve a power gain of up to 6 dB without loss in spectral efficiency. The convolutionally encoded methods [e.g., the points labeled with (2,1,6) CC (convolutional code), which



Figure 5. Bit error probability of selected rate $R = \frac{1}{2}$ convolutional codes as a function of the normalized SNR. The very large code is decoded sequentially, while the performance of the other codes, except the Turbo code, is for maximum-likelihood decoding, discussed in Section 5. (*Sources:* Refs. 8 and 9).

cies achieved by various coded and uncoded transmission methods.

Figure 6. Spectral and power efficien-

is a rate $R = \frac{1}{2}$ convolutional code with 2^6 states, and (4,1,14) CC, a rate $R = \frac{1}{4}$ convolutional code with 2^{14} states] achieve a gain in power efficiency, at the expense of spectral efficiency.

2.3. A Brief History of Error Control Coding

Trellis coding celebrated its first success in the application of convolutional codes to deep-space probes in the 1960s and 1970s. For a long time afterward, error-control coding was considered a curiosity with deep-space communications as its only viable application.

If we start with uncoded binary phase shift keying (BPSK) as our baseline transmission method, and assume coherent detection, we can achieve a bit error rate of $P_b = 10^{-5}$ at a bit energy : noise power spectral density ratio of $E_b/N_0 = 9.6$ dB, and a spectral efficiency of 1 bit per dimension. From the Shannon limit in Fig. 6 it can be seen that error-free transmission is theoretically

achievable with $E_b/N_0 = -1.59$ dB, indicating that a power savings of over 11 dB is possible by using FEC.

One of the early attempts to close this signal energy gap was the use of a rate $\frac{6}{32}$ biorthogonal (Reed–Muller) block code [4]. This code was used on the *Mariner*, *Mars* and *Viking* missions. This system had a spectral efficiency of 0.1875 bits/symbol and an optimal soft-decision decoder achieved $P_b = 10^{-5}$ with an $E_b/N_0 = 6.4$ dB. Thus, the (32,6) biorthogonal code required 3.2 dB less power than BPSK at the cost of a fivefold increase in the bandwidth (see Fig. 6).

In 1967, a new algebraic decoding technique was discovered for Bose–Chaudhuri–Hocquenghem (BCH) codes, which enabled the efficient *hard-decision decoding* of an entire class of block codes. For example, the (255,223) BCH code has an $\eta \approx 0.5$ bits/symbol, assuming ideal pulse shaping, and achieves $P_b = 10^{-5}$ with $E_b/N_0 = 5.7$ dB using algebraic decoding.

Sequential decoding allowed the decoding of longconstraint-length convolutional codes, and was first used on the *Pioneer* 9 mission. The *Pioneer* 10 and 11 missions in 1972 and 1973 both used a long-constraintlength (2,1,31) nonsystematic convolutional code [26]. A sequential decoder was used that achieved $P_b = 10^{-5}$ with $E_b/N_0 = 2.5$ dB, and $\eta = 0.5$. This is only 2.5 dB away from the capacity of the channel.

The Voyager spacecraft launched in 1977 used a shortconstraint-length (2,1,6) convolutional code in conjunction with a soft-decision optimal decoder achieving $P_b = 10^{-5}$ at $E_b/N_0 = 4.5$ dB and a spectral efficiency of $\eta = 0.5$ bits/symbol. The biggest such optimal decoder built to date [27] found application in the *Galileo* mission, where a (4,1,14) convolutional code is used. This code has a spectral efficiency of $\eta = 0.25$ bits/symbol and achieves $P_b = 10^{-5}$ at $E_b/N_0 = 1.75$ dB. Its performance is therefore also 2.5 dB away from the capacity limit. The systems for *Voyager* and *Galileo* are further enhanced by the use of concatenation in addition to the convolutional inner code. An outer (255,223) Reed–Solomon code [4] is used to reduce the required signal-to-noise ratio by 2.0 dB for the *Voyager* system and by 0.8 dB for the *Galileo* system.

More recently, Turbo codes [6] using iterative decoding virtually closed the gap to capacity by achieving $P_b = 10^{-5}$ at a spectacularly low E_b/N_0 of 0.7 dB with $R_d = 0.5$ bits/symbol. It appears that the half-century of effort to reach capacity has been achieved with this latest invention. More recently, low-density parity-check codes have also been demonstrated to obtain performances very close to capacity.

Space applications of error-control coding have met with spectacular success, and were for a long time the major, if not only, area of application for FEC. The belief that coding was useful only in improving power efficiency of digital transmission was prevalent. This attitude was overturned only by the spectacular success of error-control coding on voiceband data transmission modems. Here it was not the power efficiency that was the issue, but rather the spectral efficiency, that is given a standard telephone channel with an essentially fixed bandwidth and SNR, what was the maximum practical rate of reliable transmission.

The first commercially available voice-band modem in 1962 achieved a transmission rate of 2400 bps. Over the next 10–15 years these rates improved to 9600 bps, which was then considered to be the maximum achievable rate, and efforts to push the rate higher were frustrated. Ungerböck's invention of trellis-coded modulation in the late 1970s, however, opened the door to further, unanticipated improvements. The modem rates jumped to 14,400 bps and then to 19,200 bps, using sophisticated TCM schemes [28]. The latest chapter in voiceband data modems is the establishment of the CCITT (Consultative Committee for International Telephony and Telegraphy) V.34 modem standard [29]. The modems specified therein achieve a maximum transmission rate of 28,800 bps, and extensions to V.34 (V.34bis) to cover two new rates at 31,200 bps and 33,600 bps have been established. These rates need to be compared to estimates of the channel capacity for a voiceband telephone channel, which are somewhere around 30,000 bps, essentially achieving capacity on this channel also. Note that the common 56kbps modems used in the downward direction exploit the higher capacity provided by a digital connection to the switching station.

3. TRELLIS CODING

A trellis encoder consists of two parts, a finite-state machine (FSM) which generates the trellis of the code, and a modulator, called the signal mapper, which maps state transitions of the FSM into output symbols suitable for transmission. This encoder/modulator pair is shown in Fig. 7 for a specific example. The FSM in this example has a total of eight states, where the state s_r at time r of the FSM is defined by the contents of the delay cells: $s_r = (s_r^{(2)}, s_r^{(1)}, s_r^{(0)})$. The signal mapper performs a memoryless mapping of the 3 bits $v_r = (u_r^{(2)}, u_r^{(1)}, v_r^{(0)})$ into one of the eight symbols of an 8-PSK signal set. The FSM accepts 2 input bits $u_r = (u_r^{(2)}, u_r^{(1)}), u_r^{(i)} = \{0, 1\}$ at each symbol time r, and transitions from a state s_r to one of four possible successor states s_{r+1} . In this fashion the trellis encoder generates the (possibly) infinite sequence of symbols $\underline{x} = (\dots, x_{-1}, x_0, x_1, x_2, \dots)$. There are four choices at each time r, which allows us to transmit 2 information bits/symbol, the same as for QPSK, but using a larger constellation. This fact is called signal set expansion, and is necessary in order to introduce the redundancy required for error control.

A graphical interpretation of the function of the FSM is given by the state-transition diagram shown in Fig. 8. The nodes in this transition diagram are the possible states of the FSM, and the branches represent the possible transitions between them. Each branch can now be labeled by the pair of input bits $u = (u^{(2)}, u^{(1)})$ which cause the transition, and by either the output triple $v = (u^{(2)}, u^{(1)}, v^{(0)})$, or the output signal x(v). [In Fig. 8 we have used x(v), represented in octal notation, i.e., $x_{oct}(v) = u^{(2)}2^2 + u^{(1)}2^1 + v^{(0)}2^0$.]

If we index the states by both their content and the time index r, Fig. 8 expands into the *trellis diagram*, or simply the *trellis* of the code, shown in Fig. 9. This trellis is the two-dimensional representation of the state



Figure 7. Trellis encoder with an 8-state finite-state machine (FSM) driving a 3-bit to 8-PSK signal mapper. All inputs and outputs of the FSM and all operations are binary modulo-2 operations.

Figure 8. State-transition diagram of the encoder from Fig. 7. The labels on the branches are the encoder output signals x, in decimal notation.



Figure 9. Section of the trellis of the encoder in Fig. 7. The two solid lines depict two possible paths with their associated signal sequences through this trellis. The numbers on top are the signals transmitted if the encoder follows the upper path, and the numbers at the bottom are those on the lower path.

and time-space of the encoder; it captures all achievable states at all time intervals, usually starting from an originating state (commonly state 0), and terminating in a final state (commonly also state 0). In practice the length of the trellis will be several hundred or thousands of time units, possibly even infinite, corresponding to continuous operation. When and where to terminate the trellis is a matter of practical consideration.

Each path through the trellis corresponds to a unique message, as a sequence of symbols, and is associated with a unique sequence of signals. The term *trellis-coded modulation* originates from the fact that these encoded sequences consist of high-level modulated symbols, rather than simple binary symbols.

The FSM puts restrictions on the symbols that can be in a sequence, and these restrictions are exploited by a smart decoder. In fact, what counts is the distance between signal sequences *x*, and not the distance between individual signals as in uncoded transmission. Let us then assume that such a decoder can follow all possible sequences through the trellis, and it makes decisions between sequences. This is illustrated in Fig. 9 for two sequences $\underline{x}^{(e)}$ (erroneous) and $\underline{x}^{(c)}$ (correct). These two sequences differ in the three symbols shown. An optimal decoder will make an error between these two sequences with probability $P_s = Q(\sqrt{d_{ec}^2 E_s/2N_0})$, where $d_{ec}^2 = 4.586 =$ 2 + 0.586 + 2 is the squared Euclidean distance between $\underline{x}^{(e)}$ and $\underline{x}^{(c)}$, which is much larger than the QPSK distance of $d^2 = 2$, and E_s is the energy per signal. Examining all possible sequence pairs $\underline{x}^{(e)}$ and $\underline{x}^{(c)}$, one finds that those highlighted in Fig. 9 have the smallest squared Euclidean distance, and hence, the probability that the decoder makes an error between those two sequences is the most likely error event.

We now see that by virtue of performing sequence decoding, rather than symbol decoding, the distances between competing candidates can be increased, even though the signal constellation used for sequence coding has a smaller minimum distance between signal points than the uncoded constellation for the same rate. For this code, we may decrease the symbol power by about 3.6 dB; thus, we use less than half the power needed for QPSK to achieve the same performance. This superficial analysis belies the complexity of a more precise error analysis [5] but serves as a crude estimate.

A more precise error analysis is not quite so simple since the possible error paths in the trellis are highly correlated, which makes an exact analysis of the error probability impossible for all except the most simple cases. In fact, much work has gone into analyzing the error behavior of trellis codes [5].

4. CONSTRUCTION OF CODES

From Fig. 9 we see that an error path diverges from the correct path at some state and merges with the correct path again at a (possibly) different state. The task of designing a good trellis code means designing a trellis code for which different symbol sequences are separated by large squared Euclidean distances. Of particular importance is the minimum squared Euclidean distance, termed d_{free}^2 , namely, $d_{\text{free}}^2 = \min_{\underline{x}^{(i)}, \underline{x}^{(j)}} \|\underline{x}^{(i)} - \underline{x}^{(j)}\|^2$. A code with a large d_{free}^2 is generally expected to perform well, and d_{free}^2 has become the major design criterion for trellis codes.

One heuristic design rule [11], which was used successfully in designing codes with large d_{free}^2 , is based on the following observation. If we assign to the branches leaving a state signals from a subset with large distances between points, and likewise assign such signals to the branches merging into a state, we are assured that the total distance is at least the sum of the minimum distances between the signals in these subsets. For our 8-PSK code example, we can choose these subsets to be QPSK signal subsets of the original 8-PSK signal set. This is done by partitioning the 8-PSK signal set into two QPSK sets as illustrated in Fig. 10. The mapper function is now chosen such that the state information bit $v^{(0)}$ selects the subset and the input bits u select a signal within the subset. Since all branches leaving a state have the same state information bit $v^{(0)}$, all the branch signals are in either subset A or subset B, and the difference between two signal sequences picks up an incremental distance of $d^2 = 2$ over the first branch of their difference. These are Ungerböeck's [11] original design rules.

The values of the tap coefficients (see Fig. 13) $h^{(2)} = 0, h_1^{(2)}, \ldots, h_{\nu-1}^{(2)}, 0; h^{(1)} = 0, h_1^{(1)}, \ldots, h_{\nu-1}^{(1)}, 0;$ and $h^{(0)} = 0, h_1^{(0)}, \ldots, h_{\nu-1}^{(0)}, 0$ in the encoder are usually found via computer search programs or heuristic construction algorithms. The parameter ν is called the *constraint length*, and is the length of the shortest error path. Table 1 shows the best 8-PSK trellis codes found to date using 8-PSK with *natural mapping*, which is the bit assignment shown in Fig. 7. The figure gives the connector coefficients, $d_{\rm free}^2$, the average number $A_{d_{\rm free}}$ of paths with $d_{\rm free}^2$, and the average number $B_{d_{\rm free}}$, as well as the higher-order average path pair number, called *multiplicities*, are important parameters determining the error performance of a trellis code. The tap coefficients are given in octal notation, for instance, $h^{(0)} = 23 = 10111$, where a 1 means connected and a 0 means no connection.

From Table 1 one can see that an asymptotic coding gain (coding gain for $SNR \rightarrow \infty$ over the reference constellation that is used for uncoded transmission at the same rate) of about 6 dB can quickly be achieved



Figure 10. 8-PSK signal set partitioned into constituent QPSK signal sets.

Number of States	$h^{(0)}$	$h^{(1)}$	$h^{(2)}$	$d_{ m free}^2$	$A_{d_{\mathrm{free}}}$	$B_{d_{\mathrm{free}}}$	Asymptotic Coding Gain (dB)
4	5	2	_	4.00^{a}	1	1	3.0
8	11	2	4	4.59^{a}	2	7	3.6
16	23	4	16	5.17^{a}	2.25	11.5	4.1
32	45	16	34	5.76^{a}	4	22.25	4.6
64	103	30	66	6.34^{a}	5.25	31.125	5.0
128	277	54	122	6.59^{a}	0.5	2.5	5.2
256	435	72	130	7.52^{a}	1.5	12.25	5.8
512	1,525	462	360	7.52^{a}	0.313	2.75	5.8
1,024	2,701	1,216	574	8.10^{a}	1.32	10.563	6.1
2,048	4,041	1,212	330	8.34	3.875	21.25	6.2
4,096	15,201	6,306	4,112	8.68	1.406	11.758	6.4
8,192	20,201	12,746	304	8.68	0.617	2.711	6.4
32,768	$143,\!373$	70,002	47,674	9.51	0.25	2.5	6.8
131,072	616,273	340,602	$237,\!374$	9.85	_	_	6.9

 Table 1. Connectors, Free-Squared Euclidean Distance, and Asymptotic

 Coding Gains of Some Maximum Free-Distance 8-PSK Trellis Codes

^aCodes found by exhaustive computer searches [13,33]; other codes (without the^a superscript) were found by various heuristic search and construction methods [33,34]. The connector polynomials are in octal notation.

with moderate effort. Since the asymptotic coding gain is a reasonable yardstick at the bit error rates of interest, codes with a maximum of about 1000 states seem to exploit most of what can be gained by this type of coding.

Some other researchers have used different mapper functions in an effort to improve performance, in particular the bit error performance that could be improved by up to 0.5 dB. (For instance, 8-PSK Gray mapping [labeling the 8-PSK symbols successively by (000), (001), (011), (010), (110), (111), (101), (100)] was used by Du and Kasahara [31] and Zhang [32]. Zhang also used another mapper [labeling the symbols successively by (000), (001), (010), (011), (110), (111), (100), (101)] to further improve on the bit error multiplicity. The search criterion employed involved minimizing the bit multiplicities of several spectral lines in the distance spectrum of a code. Table 2 gives the best 8-PSK codes found so far with respect to the bit error probability.

If we go to higher-order signal sets such as 16-QAM, 32cross, and 64-QAM, there are, at some point, not enough states left such that each diverging branch leads to a different state, and we have parallel transitions, that is, two or more branches connecting two states. Naturally we would want to assign signals with large distances to such parallel branches to avoid a high probability of error,

Table 2. Table of Improved 8-PSK Codes Using aDifferent Mapping Function [32]

Number of States	$h^{(0)}$	$h^{(1)}$	$h^{(2)}$	$d_{ m free}^2$	$A_{d_{\mathrm{free}}}$	$B_{d_{\mathrm{free}}}$
8	17	2	6	4.59	2	5
16	27	4	12	5.17	2.25	7.5
32	43	4	24	5.76	2.375	7.375
64	147	12	66	6.34	3.25	14.8755
128	277	54	176	6.59	0.5	2
256	435	72	142	7.52	1.5	7.813
512	1377	304	350	7.52	0.0313	0.25
1024	2077	630	1132	8.10	0.2813	1.688

since the probability of these errors cannot be influenced by the code.

The situation of parallel transition is actually the case for the first 8-PSK code in Table 1, whose trellis is given in Fig. 11. Here the parallel transitions are by choice, not by necessity. Note that the minimum-distance path pair through the trellis has $d^2 = 4.586$, but that is not the most likely error to happen. All signals on parallel branches are from a BPSK subset of the original 8-PSK set, and hence their distance is $d^2 = 4$, which gives the 3-dB asymptotic coding gain of the code over QPSK.

In general, we partition a signal set into a partition chain of subsets, such that the minimum distance between signal points in the new subsets is maximized at every level. This is illustrated in Fig. 12 with the 16-QAM signal set and a binary partition chain, which splits each set into two subsets at each level. Note that the partitioning can be continued until there is only one signal left in each subset. In such a way, by following the partition path, a "natural" binary label can be assigned to each signal point. This method of partitioning a signal set is called *set partitioning* with increasing intrasubset distances.

The idea is to use these constellations for codes with parallel transitions by not encoding all the input bits of u_r . Using the encoder in Fig. 13 with a 16-QAM constellation,



Figure 11. Four-state 8-PSK trellis code with parallel transitions.



Figure 12. Set partitioning of a 16-QAM signal set into subsets with increasing minimum distance. The final partition level used by the encoder in Fig. 13 is the fourth level, that is the subsets with two signal points each.



Figure 13. Generic encoder for QAM signal constellations.

for example, the first information bit $u_r^{(3)}$ is not encoded and the output signal of the FSM selects now a subset rather than a signal point. This subset is at the fourth partition level in Fig. 12. The uncoded bit(s) select the actual signal point within the subset. Analogously, then, the encoder now has to be designed to maximize the minimum interset distances of sequences, since it cannot influence the signal point selection within the subsets. The advantage of this strategy is that the same encoder can be used for all signal constellations with the same intraset distances at the final partition level, in particular for all signal constellations that are nested versions of each other, such as 16-QAM, 32-cross, and 64-QAM.

Figure 13 shows such a generic encoder that maximizes the minimum interset distance between sequences, and it can be used with all QAM-based signal constellations. Only the two least significant information bits affect the encoder FSM. All other information bits cause parallel transitions. Table 3 shows the coding gains achievable with such an encoder structure. The gains when going from 8-PSK to 16-QAM are most marked since rectangular constellations have a better power efficiency than do constant-energy circular constellations.

As in the case of 8-PSK codes, efforts have been made to improve the distance spectrum of a code, in particular to minimize the bit error multiplicity of the first few spectral lines. Some of the improved codes using a 16-QAM constellation are listed in Table 4 together with the original codes from Table 3, which are marked by "Ung." The improved codes, taken from Zhang [32], use the signal mapping shown in Fig. 14. Note also that the input line $u_r^{(3)}$ is also fed into the encoder for these codes.

	Connectors				Asymptotic Coding Gain (dB)		
Number of States	$h^{(0)}$	$h^{(1)}$	$h^{(2)}$	$d_{ m free}^2$	16-QAM/ 8-PSK	32-cross/ 16-QAM	64-QAM 32-cross
4	5	2	_	4.0	4.4	3.0	2.8
8	11	2	4	5.0	5.3	4.0	3.8
16	23	4	16	6.0	6.1	4.8	4.6
32	41	6	10	6.0	6.1	4.8	4.6
64	101	16	64	7.0	6.8	5.4	5.2
128	203	14	42	8.0	7.4	6.0	5.8
256	401	56	304	8.0	7.4	6.0	5.8
512	1001	346	510	8.0	7.4	6.0	5.8

 Table 3. Connectors and Gains of Maximum Free Distance QAM

 Trellis Codes

Source: The codes in this table were presented by Ungerböeck [13].

 Table 4. Original 16-QAM Trellis Code and Improved

 Trellis Codes Using Nonstandard Mapping

2^{ν}	$h^{(0)}$	$h^{(1)}$	$h^{(2)}$	$h^{(3)}$	$d_{ m free}^2$	$A_{d_{\mathrm{free}}}$	$B_{d_{\mathrm{free}}}$
Ung 8	11	2	4	0	5.0	3.656	18.313
8	13	4	2	6	5.0	3.656	12.344
Ung 16	23	4	16	0	6.0	9.156	53.5
16	25	12	6	14	6.0	9.156	37.594
Ung 32	41	6	10	0	6.0	2.641	16.063
32	47	22	16	34	6.0	2	6
Ung 64	101	16	64	0	7.0	8.422	55.688
64	117	26	74	52	7.0	5.078	21.688
Ung 128	203	14	42	0	8.0	36.16	277.367
128	313	176	154	22	8.0	20.328	100.031
Ung 256	401	56	304	0	8.0	7.613	51.953
256	417	266	40	226	8.0	3.273	16.391

5. CONVOLUTIONAL CODES

Convolutional codes historically were the first trellis codes. They were introduced in 1955 by Elias [54]. Since then, much theory has evolved to understand convolutional codes. A convolutional code is obtained by using a special modulator. The output binary digits of the encoder, shown again in Fig. 15, are no longer jointly encoded into a modulation symbol, but each bit is encoded into a binary BPSK signal, using the mapping $0 \rightarrow -1, 1 \rightarrow 1$.

It is immediately clear now that the input and output symbol rates are different. For example, in Fig. 15, this rate changes from 2 input bits/time unit to 3 output symbols/time unit. This causes a bandwidth expansion of $\frac{3}{2}$ with respect to uncoded BPSK modulation. It is this bandwidth expansion that is partly responsible for the excellent performance of convolutional codes.

It is interesting to note that the convolutional encoder in Fig. 15 has an alternate "incarnation," which is given in Fig. 16. This form is called the *controller canonical nonsystematic form*; the term stems from the fact that inputs can be used to control the state of the encoder in a very direct way, in which the outputs have no influence. Both encoders generate the same code, but individual input bit sequences map onto different output bit sequences.

The squared Euclidean distance between two signal sequences depends only on the Hamming distance $H_d(\underline{v}^{(1)}, \underline{v}^{(2)})$ between the two output symbol sequences, where the Hamming distance between two sequences is defined as the number of bit positions in which the two sequences differ. Consequently, the minimum squared Euclidean distance d_{free}^2 depends only on the number of binary differences between the closest code sequences. This number is the minimum Hamming distance of a convolutional code, denoted by d_{free} . Convolutional codes are linear in the sense that the modulo-2 addition of



Figure 14. Mapping used for the improved 16-QAM codes.



Figure 15. Rate $R = \frac{2}{3}$ convolutional code which was used in Fig. 7 to generate an 8-PSK trellis code.



Figure 16. Rate $R = \frac{2}{3}$ convolutional code from above in controller canonical non-systematic form.

two output bit sequences is another valid code sequence, and finding the minimum Hamming distance between two sequences $\underline{v}^{(1)}$ and $\underline{v}^{(2)}$ amounts to finding the minimum Hamming weight of any code sequence \underline{v} .

Finding convolutional codes with large minimum Hamming distance is exactly as difficult as finding good general trellis codes with large Euclidean free distance, and, as in the case for trellis codes, computer searches are usually used to find good codes [58–60]. Most often the controller canonical form (Fig. 16) of an encoder is preferred in these searches. The procedure is then to search for a code with the largest minimum Hamming weight by varying the taps either exhaustively or according to heuristic rules. In this fashion, the codes in Tables 5–12 were found [15–17,59]. They are the rate $R = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}$, and $R = \frac{2}{3}$ codes with the greatest minimum Hamming distance $d_{\rm free}$ for a given constraint length.

6. DECODING

6.1. Sequence Decoding

There are basically two major decoding strategies in common use. These are the sequence decoders and the

Table 5. Connectors^{*a*} and Free Hamming Distance of the Best $R = \frac{1}{2}$ Convolutional Codes [16]

-			
Constraint Length v	$g^{(1)}$	$g^{(0)}$	d_{free}
2	5	7	5
3	15	17	6
4	23	35	7
5	65	57	8
6	133	171	10
7	345	237	10
8	561	753	12
9	1161	1545	12
10	2335	3661	14
11	4335	5723	15
12	10533	17661	16
13	21675	27123	16
14	56721	61713	18
15	111653	145665	19
16	347241	246277	20

^{*a*}The connectors are given in octal notation (e.g., g = 17 = 1111).

Table 6. Connectors and Free Hamming Distance of the Best $R = \frac{1}{3}$ Convolutional Codes [16]

Constraint Length v	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
9	5	7	7	8
2	19	15	17	10
3	10	10	17	10
4	25	33	37	12
5	47	53	75	13
6	133	145	175	15
7	225	331	367	16
8	557	663	711	18
9	1,117	1,365	1,633	20
10	2,353	2,671	3,175	22
11	4,767	5,723	6,265	24
12	10,533	10,675	17,661	24
13	$21,\!645$	35,661	37,133	26

Table 7. Connectors and Free Hamming
Distance of the Best $R = \frac{2}{3}$ Convolutional
Codes [16]

Constraint Length v	$g_2^{(2)}, g_1^{(2)}$	$g_2^{(1)}, g_1^{(1)}$	$g_2^{(0)}, g_1^{(0)}$	d_{free}
2	3,1	1,2	3,2	3
3	2,1	1,4	3,7	4
4	7,2	1,5	4,7	5
5	14,3	6,10	16,17	6
6	15,6	6,15	15,17	7
7	14,3	7,11	13,17	8
8	32,13	5,33	25,22	8
9	25,5	3,70	36,53	9
10	63,32	15,65	46,61	10

symbol decoders. A sequence decoder looks for the most likely transmitted code sequence. Thus, if \underline{y} is the received sequence from the channel, an optimal sequence decoder calculates the conditional probability

$$\underline{\hat{x}} = \max_{\underline{x}} \Pr(\underline{x}|\underline{y}) = \max_{\underline{x}} \Pr(\underline{y}|\underline{x}), \quad (8)$$

assuming that all \underline{x} are a priori equally likely to have been sent. Such a decoder is referred to as a *maximum-likelihood* (ML) *sequence decoder*. For a trellis code, this amounts to an algorithm which exhaustively

Table 8. Connectors and Free Hamming Distance of the Best $R = \frac{1}{4}$ Convolutional Codes [58]

Constraint Length v	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	5	7	7	7	10
3	13	15	15	17	13
4	25	27	33	37	16
5	53	67	71	75	18
6	135	135	147	163	20
7	235	275	313	357	22
8	463	535	733	745	24
9	1,117	1,365	1,633	1,653	27
10	2,387	2,353	2,671	3,175	29
11	4,767	5,723	6,265	$7,\!455$	32
12	11,145	12,477	15,537	16,727	33
13	21,113	$23,\!175$	35,527	35,537	36

Table 9. Connectors and Free Hamming Distance of the Best $R = \frac{1}{5}$ Convolutional Codes [15]

Constraint Length v	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	g ⁽⁰⁾	d_{free}
2	7	7	7	5	5	13
3	17	17	13	15	15	16
4	37	27	33	25	35	20
5	75	71	73	65	57	22
6	175	131	135	135	147	25
7	257	233	323	271	357	28

Table 10. Connectors and Free Hamming Distance of the Best $R = \frac{1}{6}$ Convolutional Codes [15]

Constraint Length v	$g^{(5)}$	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	7	7	7	7	5	5	16
3	17	17	13	13	15	15	20
4	37	35	27	33	25	35	24
5	73	75	55	65	47	57	27
6	173	151	135	135	163	137	30
7	253	375	331	235	313	357	34

searches through all the possible paths through the trellis. However, some simplifications are possible.

The algorithms will start in the known state at the beginning of the trellis, and explore all possible paths

one step at a time. It will keep a measure of reliability at each state and time, called the *state metric*. This state metric for state $s_n = i$ at time *n* is calculated as the partial path probability $\Pr(\underline{\tilde{y}}|\underline{\tilde{x}})$ for the partial path $\underline{\tilde{x}}(i) = (\ldots, x_0, x_1, \ldots, x_n)$ that leads to state $s_n = i$, given the partial received sequence $\underline{\tilde{y}} = (\ldots, y_0, y_1, \ldots, y_n)$ up to time *n*.

This probability can be expressed recursively, and the metric at state $s_n = i$ and time n, denoted by $J_n(i)$, can be calculated from previous state metrics according to

$$J_n(i) = J_{n-1}(j) + \lambda_n(j \to i) \tag{9}$$

In other words, the metric at state *i* at time *n* is calculated from the metric at state *j* at time n-1 by adding a measure $\lambda_n(j \rightarrow i)$, called the *branch metric*, that depends on the symbol on the transition from $j \rightarrow i$ and the received signal y_n . State *j* must connect to state *i* in the trellis of the code. However, state *j* is not usually the only state that connects to state *i*, and the algorithm will select only the best metric among all merging connections. This is known as the *add-compare-select* step in the algorithm and is formally given by

$$J_n(i) = \min_{j \to i} (J_{n-1}(j) + \lambda_n(j \to i))$$
(10)

The algorithm furthermore needs to remember the history of selection decisions, since the winning sequence can be determined only at the end of the received sequence. This method was introduced by Viterbi in 1967 [18,19] in the context of analyzing convolutional codes, and has since become widely known as the *Viterbi algorithm* [20]. The algorithm for a block of length L is as follows:

Step 1. Initialize the S states of the maximumlikelihood decoder with the metric $J_0(i) = -\infty$ and survivors $\underline{\tilde{x}}(i) = \{ \}$. Initialize the starting state of the decoder, usually state i = 0, with the metric $J_0(0) = 0$. Let n = 1.

Step 2. Calculate the branch metric

$$\lambda_n = |y_n - x_n(j \to i)|^2 \tag{11}$$

for each state j and each signal $x_n(j \rightarrow i)$ that is attached to the transition from state i to state j.

Step 3. For each state *i*, choose from the 2^k merging paths the survivor $\underline{\tilde{x}}(i)$ for which $J_n(i)$ is maximized.

Table 11. Connectors and Free Hamming Distance of theBest $R = \frac{1}{7}$ Convolutional Codes [15]

Constraint Length v	$g^{(6)}$	$g^{(5)}$	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	7	7	7	7	5	5	5	18
3	17	17	13	13	13	15	15	23
4	35	27	25	27	33	35	37	28
5	53	75	65	75	47	67	57	32
6	165	145	173	135	135	147	137	36
7	275	253	375	331	235	313	357	40

Table 12. Connectors and Free Hamming Distance of the Best $R = \frac{1}{8}$ Convolutional Codes [15]

Constraint Length v	g ⁽⁷⁾	$g^{(6)}$	$g^{(5)}$	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	7	7	5	5	5	7	7	7	21
3	17	17	13	13	13	15	15	17	26
4	37	33	25	25	35	33	27	37	32
5	57	73	51	65	75	47	67	57	36
6	153	111	165	173	135	135	147	137	40
7	275	275	253	371	331	235	313	357	45

- Step 4. If n < L, let n = n + 1 and go to step 2, or else go to step 5.
- Step 5. Output the survivor $\underline{x}(i)$ that maximizes $J_L(i)$ as the maximum-likelihood estimate of the transmitted sequence.

The Viterbi algorithm has enjoyed tremendous popularity, not only in decoding trellis codes but also in symbol sequence estimation over channels affected by intersymbol interference [17,21] and multiuser optimal detectors [22]. Whenever the underlying generating process can be modeled as a finite-state machine, the Viterbi algorithm finds application.

A rather large body of literature deals with the Viterbi decoder, and there are a number of books dealing with the subject [e.g., 16,17,23,24]. One of the more important results is that it can be shown that there is no need to wait until the entire sequence is decoded before starting to output the estimated symbols \tilde{x}_n , or the corresponding data. The probability that the symbols in all survivors $\underline{\tilde{x}}(i)$ are identical for $m \leq n - n_t$ is very close to unity for $n_t \approx 5\nu$. n_t is called the *truncation length* or *decision depth*. We may therefore modify the algorithm to obtain a fixed-delay decoder by modifying steps 4 and 5 of the algorithm outlined above as follows:

- Step 4. If $n \ge n_t$, output $x_{n-n_t}(i)$ from the survivor $\underline{\tilde{x}}(i)$ with the largest metric $J_n(i)$ as the estimated symbol at time $n n_t$. If n < L, let n = n + 1 and go to step 2. Step 5. Output the remaining estimated symbols
- $x_n(i); L n_t < n \le L$ from the survivor $\underline{x}(i)$ that maximizes $J_L(i)$.

We recognize that we may now let $L \to \infty$; thus, the complexity of our decoder is not determined by the length of the sequence, and it may be operated in a continuous fashion.

6.2. Symbol Decoding

The other important class of decoders targets symbols, rather than sequences. The goal is to calculate the a posteriori probability (APP)

$$\Pr(u_r|y) \tag{12}$$

that is, the probability of u_r after observing <u>y</u>. This value can be used to find the maximum APP (MAP) decision of u_r as

$$\hat{u}_r = \max_{u_r} \Pr(u_r | \underline{y}) \tag{13}$$

However, as a symbol estimator, the resulting error probability is not significantly better than that achieved with the Viterbi decoder (8). The soft value of (12) is used mostly in iterative decoding such as Turbo decoding, where soft output values from component decoders are required.

The APP of u_r is calculated by the *backward-forward* algorithm, a procedure that sweeps through the trellis in both the forward and the backward directions. The algorithm functions as follows. First the probability of the transition from state j to i, given \underline{y} is calculated. It can be broken into three factors, given by

$$\Pr(s_{r-1} = j, s_r = i, y) = \alpha_{r-1}(j)\gamma_r(j \to i)\beta_r(i)$$
(14)

where the α values are the result of the forward pass and are calculated recursively according to

$$\alpha_r(j) = \sum_{\text{states } l} \alpha_{r-1}(l) \gamma_r(l \to j)$$
(15)

Furthermore, for a trellis code started in the zero state at time r = 0 we have the starting conditions

$$\alpha_0(0) = 1, \, \alpha_0(j) = 0; \quad j \neq 0 \tag{16}$$

Similarly

$$\beta_r(i) = \sum_{\text{states } l} \beta_{r+1}(l) \gamma_{r+1}(i \to l) \tag{17}$$

The boundary condition for $\beta_r(i)$ for a code that is terminated in the zero state at time r = L is

$$\beta_L(0) = 1, \, \beta_L(i) = 0; \quad i \neq 0$$
(18)

The values $\gamma_r(j \to i)$ associated with the transition from state *j* to state *i* are external values and are calculated as

$$\gamma_r(j \to i) = \sum_{x_r} p_{ij} q_{ij}(x_r) p_n(y_r - x_r)$$
(19)

where $p_n(\cdot)$ is the probability density function of the AWGN (additive white Gaussian noise) channel given by $\Pr(y_r|x_r) = p_n(y_r - x_r)$. p_{ij} is the a priori probability of the state transitions, usually equal for all transitions, and $q_{ij}(x_r)$ is the probability of choosing x_r if there is more than one signal choice per transition. The calculation of $\gamma_r(j \to i)$ is not very complex and can most easily be implemented by a table lookup procedure.

Equations (15) and (17) are iterative updates of internal variables. The complete algorithm to calculate the a posteriori state-transition probabilities is as follows:

- Step 1. Initialize $\alpha_0(0) = 1$, $\alpha_0(j) = 0$ for all non-zero states $(j \neq 0)$, and $\beta_L(0) = 1$, $\beta_L(j) = 0$, $j \neq 0$. Let r = 1.
- Step 2. Calculate $\gamma_r(j \to i)$ using (19), and $\alpha_r(j)$ using (15) for all states j.
- Step 3. If r < L, let r = r + 1 and go to step 2, else r = L 1 and go to step 4.
- Step 4. Calculate $\beta_r(i)$ using (17), and $\Pr(s_{r-1} = j, s_r = i, y)$ from (14).
- Step 5. If r > 1, let r = r 1 and go to step 4.
- Step 6. Terminate the algorithm and output all the values and $Pr(s_{r-1} = j, s_r = i, y)$.

Contrary to the ML algorithm, the MAP algorithm needs to go through the trellis twice, once in the forward direction, and once in the reverse direction. What weighs even more is that all the values $\alpha_r(j)$ must be stored from the first pass through the trellis. For a rate k/n convolutional code, for example, this requires $2^{k\nu}L$ storage locations since there are $2^{k\nu}$ states, for each of which we need to store a different value $\alpha_r(j)$ at each time epoch r. The storage requirement grows exponentially in the constraint length ν and linearly in the block length L.

The a posteriori transition probabilities produced by this algorithm can now be used to calculate a posteriori information bit probabilities, that is, the probability that the information k-tuple $u_r = u$. Starting from the transition probabilities $Pr(s_{r-1} = j, s_r = i|\underline{y})$ we simply sum over all transitions $j \rightarrow i$ that are caused by $u_r = u$. Denoting these transitions by A(u), we obtain

$$\Pr(u_r = u) = \sum_{(j \to i) \in A(u)} \Pr(s_{r-1} = j, s_r = i | \underline{y})$$
(20)

Another most interesting product of the MAP decoder is the a posteriori probability of the transmitted output symbol x_r . Arguing analogously as above, and letting B(x)be the set of transitions on which the output signal x can occur, we obtain

$$\Pr(x_r = x) = \sum_{(j \to i) \in B(x)} \Pr(x|y_r) \Pr(s_{r-1} = j, s_r = i|\underline{y})$$



where the a priori probability of y_r can be calculated via

$$p(y_r) = \sum_{x'} p(y_r | x') q_{ij}(x')$$
(22)

Equation (21) can be greatly simplified if there is only one output symbol on the transition $j \rightarrow i$. In this case, the transition automatically determines the output symbol, and

$$\Pr(x_r = x) = \sum_{(j \to i) \in B(x)} \Pr(s_{r-1} = j, s_r = i | \underline{y})$$
(23)

7. PARALLEL CONCATENATED TRELLIS CODES (TURBO CODES)

7.1. Code Structure

A Turbo encoder consists of the *parallel concatenation* of two or more, usually identical, rate $\frac{1}{2}$ encoders, realized in systematic feedback form, and a pseudorandom interleaver. This encoder structure is called a parallel concatenation because the two encoders operate on the same *set* of input bits, rather than one encoding the output of the other as in serial concatenation. A block diagram of a Turbo encoder with two constituent convolutional encoders is shown in Fig. 17.

The interleaver is used to permute the input bits such that the two encoders are operating on the same set of input bits, but different input sequences. Thus, the first encoder receives the input bit u_r and produces the output pair $(u_r, v_r^{(1)})$, while the second encoder receives the input bit u'_r and produces the output pair $(u'_r, v_r^{(2)})$. The input bits are grouped into finite-length sequences whose length, N, equals the size of the interleaver. Since both the encoders are systematic and operate on the same set of input bits, it is only necessary to transmit the input bits once and the overall code has rate $\frac{1}{3}$. In order to increase the overall rate of the code to $\frac{1}{2}$, the two parity sequences $\underline{v}^{(1)}$ and $v_r^{(2)}$. We will refer to a Turbo code whose constituent encoders have parity-check polynomials h_0 and h_1 , expressed in



Figure 17. Block diagram of a Turbo encoder with two constituent encoders and an optional puncturer.

2650 TRELLIS CODING

octal notation, and whose interleaver is of length N as an (h_0,h_1,N) Turbo code.

Several salient points concerning the structure of the codewords in a Turbo code are

- 1. Because the pseudorandom interleaver permutes the input bits, the two input sequences \underline{u} and \underline{u}' are almost always different, although of the same weight, and the two encoders will (with high probability) produce parity sequences of different weights.
- 2. It is easily seen that a codeword may consist of a number of distinct detours in each encoder. Note that since the constituent encoders are realized in systematic feedback form, a nonzero bit is required to return to the all-zero state and thus all detours are associated with information sequences of weight 2 or greater. Finally, with a pseudorandom interleaver it is highly unlikely that both encoders will be returned to the all-zero state at the end of the codeword even when the last ν bits of the input sequence \underline{u} are used to force the first encoder back to the all zero state.

If neither encoder is forced to the all-zero state, that is, no tail is used, then the sequence consisting of N-1 zeros followed by a one is a valid input sequence \underline{u} to the first encoder. For some interleavers, this \underline{u} will be permuted to itself and \underline{u}' will be the same sequence. In this case, the maximum weight of the codeword with puncturing, and thus the free distance of the code, will be 2! For this reason, it is common to assume that the first encoder is forced to return to the all zero state. The ambiguity of the final state of the second encoder results in negligible performance degradation for large interleavers.

7.2. Iterative Decoding of Turbo Codes

It is clear from the discussion of the codeword structure of Turbo codes that the state space of these codes is too large to perform optimum decoding. To overcome this, the discoverers of Turbo codes proposed a novel iterative decoder based on the a posteriori (AAP) symbol decoding algorithm.

The AAP decoder for each component code computes the a posteriori probability $Pr(u_r = u | \underline{y})$ conditioned on the received sequence \underline{y} . The iterative Turbo decoder makes use of these a posteriori probabilities in the form of a log-likelihood ratio (LLR) given by

$$L(u_r) = \log \frac{\Pr(u_r = 1|y)}{\Pr(u_r = 0|y)}$$
(24)

which, from (14) and (20), is given by

$$L(u_r) = \log \frac{\sum_{\substack{(j \to i) \in A(u_r=1)}} \gamma_r(j \to i) \alpha_{r-1}(j) \beta_r(i)}{\sum_{\substack{(j \to i) \in A(u_r=0)}} \gamma_r(j \to i) \alpha_{r-1}(j) \beta_r(i)}.$$
 (25)

Let $y_r^{(0)}$ be the received systematic bit and $y_r^{(m)}$, m = 1, 2, the received parity bit corresponding to the *m*th constituent encoder. With these, $\gamma_r(j \to i)$ may be expressed as

$$\nu_r(j \to i) = p_{ij} \Pr(y_r^{(0)}, y_r^{(m)} | u_r, v_r^{(m)})$$
(26)

For systematic codes, (26) may be factored as

$$\Pr(y_r^{(0)}, y_r^{(m)} | u_r, v_r^{(m)}) = \Pr(y_r^{(0)} | u_r) \Pr(y_r^{(m)} | v_r^{(m)})$$
(27)

since the received systematic sequence and the received parity sequence are conditionally independent of each other. Finally, substituting (26) and (27) into (25) and factoring yields

$$\begin{split} L(u_r) &= \log \frac{\sum_{\substack{(j \to i) \in A(u_r=1)}} \Pr(y_r^{(m)} | v_r^{(m)}) \alpha_{r-1}(j) \beta_r(i)}{\sum_{\substack{(j \to i) \in A(u_r=0)}} \Pr(y_r^{(m)} | v_r^{(m)}) \alpha_{r-1}(j) \beta_r(i)} \\ &+ \log \frac{\Pr(u_r=1)}{\Pr(u_r=0)} \\ &+ \log \frac{\Pr(y_r^{(0)} | u_r=1)}{\Pr(y_r^{(0)} | u_r=0)} = \Lambda_{e,r}^{(m)} + \Lambda_r + \Lambda_s \quad (28) \end{split}$$

where $\Lambda_{e,r}^{(m)}$ is called the *extrinsic information* from the *m*th decoder, Λ_r is the a priori log-likelihood ratio of the systematic bit u_r and Λ_s is the log-likelihood ratio of the a posteriori probabilities of the systematic bit.

A block diagram of an iterative Turbo decoder is shown in Fig 18, where each APP decoder corresponds to a constituent code. The interleavers are identical to



Figure 18. Block diagram of a Turbo decoder with two constituent decoders.

the interleavers in the Turbo encoder and are used to reorder the sequences so that each decoder is properly synchronized. For the first iteration, the first decoder computes the log-likelihood ratio of Eq. (28) with $\Lambda_r = 0$, since u_r is equally likely to be a 0 or a 1, using the received sequences $\underline{y}^{(0)}$ and $\underline{y}^{(1)}$. The second decoder now computes the log-likelihood ratio of Eq. (28) on the basis of the received sequences $y^{(0)}$ and $y^{(2)}$ (suitably reordered).

The second decoder, however, has available an estimate of the a posteriori probability of u_r from the first decoder, namely, $L^{(1)}(u_r)$. The second decoder may consider this the a priori probability of u_r in (28) and compute

$$L^{(2)}(u_r) = \Lambda^{(2)}_{e,r} + L^{(1)} + \Lambda_s$$

= $\Lambda^{(2)}_{e,r} + \Lambda^{(1)}_{e,r} + \Lambda^{(1)}_r + \Lambda_s + \Lambda_s$ (29)

as the new LLR. Close examination of (29) reveals that by passing $L^{(1)}$, the second decoder is given $\Lambda_r^{(1)}$, the previous estimate of the a priori probability, which is unnecessary. In addition, it is seen that as the decoder continues to iterate, the LLR accumulates Λ_s and the systematic bit becomes overemphasized. In order to prevent this, the second decoder subtracts $\Lambda_r^{(1)}$ and Λ_s from the information passed from the first decoder and calculates

$$L^{(2)}(u_r) = \Lambda^{(2)}_{e,r} + \Lambda^{(1)}_{e,r} + \Lambda_s$$
(30)

What is passed between the two decoders is in fact the extrinsic information only. This process continues until a desired performance is achieved, at which point a final decision is made by comparing the final log-likelihood ratio to the threshold 0.

The extrinsic information is a reliability measure of each component decoder's estimate of the transmitted sequence on the basis of the corresponding received component parity sequence and is essentially independent of the received systematic sequence. Since each component decoder uses the received systematic sequence directly, the extrinsic information allows the decoders to share information without significant error propagation. The efficacy of this technique can be seen in Fig 19, which shows the performance of the original (37,21,65536) Turbo code as a function of the decoder iterations. It is impressive that the performance of the code with iterative decoding continues to improve up to 18 iterations (and beyond).

BIOGRAPHY

Christian Schlegel received the Dipl. El. Ing. ETH degree from the Federal Institute of Technology, Zürich, Switzerland, in 1984, and his M.S. and Ph.D. degrees in electrical engineering from the University of Notre Dame, Indiana, in 1986 and 1989. From 1988 to 1992 he was with Asea Brown Boveri, Ltd., Baden, Switzerland, from 1992–1994 with the Digital Communications Group at the University of South Australia in Adelaide, from 1994–2001 he held faculty positions at the Universities of Texas at San Antonio and Utah in Salt Lake City. In 2001 he was named iCORE Professor for High-Capacity Digital Communications at the University of Alberta, Canada.



Figure 19. Performance of the (37,21,65536) Turbo code as function of the number of decoder iterations.

2652 TRELLIS CODING

His interests are in the area of digital communications and mobile radio systems, error control coding, multiple access communications, and analog and digital system implementations. He is the author of *Trellis Coding* (IEEE 1997) and *Trellis and Turbo Coding* (Wiley/IEEE 2002). Dr. Schlegel received an NSF 1997 Career Award and a Canada Research Chair in 2001.

BIBLIOGRAPHY

- C. E. Shannon, A mathematical theory of communications, Bell Syst. Tech. J. 27: 379-423 (July 1948).
- R. W. Hamming, Error detecting and error correcting codes, Bell Syst. Tech. J. 29: 147–160 (1950).
- M. J. E. Golay, Notes on digital coding, Proc. IEEE 37: 657 (1949).
- 4. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, New York, 1988.
- 5. C. Schlegel, *Trellis Coding*, IEEE Press, Piscataway, NJ, 1997.
- C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc.* 1993 IEEE Int. Conf. Communication, Geneva, Switzerland, 1993, pp. 1064–1070.
- F.-Q. Wang and D. J. Costello, Probabilistic construction of large constraint length trellis codes for sequential decoding, *IEEE Trans. Commun.* COM-43: 2439–2448 (September 1995).
- J. A. Heller and J. M. Jacobs, Viterbi detection for satellite and space communication, *IEEE Trans. Commun. Technol.* COM-19: 835-848 (Oct. 1971).
- J. K. Omura and B. K. Levitt, Coded error probability evaluation for antijam communication systems, *IEEE Trans. Commun.* COM-30: 896-903 (May 1982).
- G. Ungerböeck, J. Hagenauer, and T. Abdel-Nabi, Coded 8-PSK experimental modem for the INTELSAT SCPC system, *Proc. ICDSC*, 7th, 1986, pp. 299–304.
- G. Ungerböeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**(1): 55-67 (Jan. 1982).
- G. Ungerböeck, Trellis-coded modulation with redundant signal sets part I: Introduction, *IEEE Commun. Mag.* 25(2): 5-11 (Feb. 1987).
- G. Ungerböeck, Trellis-coded modulation with redundant signal sets part II: State of the art, *IEEE Commun. Mag.* 25(2): 12-21 (Feb. 1987).
- 14. H. Imai et al., Essentials of Error-Control Coding Techniques, Academic Press, New York, 1990.
- D. G. Daut, J. W. Modestino, and L. D. Wismer, New short constraint length convolutional code construction for selected rational rates, *IEEE Trans. Inform. Theory* **IT-28**: 793–799 (Sept. 1982).
- S. Lin and D. J. Costello, Jr., Error Control Coding: Fundamentals and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1983.
- J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
- A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**: 260–269 (April 1969).

- J. K. Omura, On the Viterbi decoding algorithm, *IEEE Trans.* Inform. Theory **IT-15**: 177–179 (Jan. 1969).
- G. D. Forney, Jr., The Viterbi algorithm, Proc. IEEE 61: 268-278 (1973).
- G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363-378 (May 1972).
- S. Verdú, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **IT-32**: 85–96 (Jan. 1986).
- 23. R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, Reading, MA, 1987.
- 24. A. J. Viterbi and J. K. Omura, *Principles of Digital Commu*nication and Coding, McGraw-Hill, New York, 1979.
- H. Nyquist, Certain topics in telegraph transmission theory, AIEE Trans. 617 ff. (1946).
- J. L. Massey and D. J. Costello, Jr., Nonsystematic convolutional codes for sequential decoding in space applications, *IEEE Trans. Commun. Technol.* COM-19: 806-813 (1971).
- O. M. Collins, The subtleties and intricacies of building a constraint length 15 convolutional decoder, *IEEE Trans. Commun.* COM-40: 1810–1819 (1992).
- G. D. Forney, Jr., Coded modulation for bandlimited channels, *IEEE Inform. Theory Soc. Newsl.* (Dec. 1990).
- 29. CCITT Recommendations V.34.
- U. Black, The V Series Recommendations, Protocols for Data Communications over the Telephone Network, McGraw-Hill, New York, 1991.
- J. Du and M. Kasahara, Improvements of the information-bit error rate of trellis code modulation systems, *IEICE*, *Japan* E72: 609-614 (May 1989).
- 32. W. Zhang, *Finite-State Machines in Communications*, Ph.D. thesis, Univ. South Australia, Australia, 1995.
- 33. J. E. Porath and T. Aulin, Fast Algorithmic Construction of Mostly Optimal Trellis Codes, Technical Report 5, Division of Information Theory, School of Electrical and Computer Engineering, Chalmers Univ. Technology, Göteborg, Sweden, 1987.
- J. E. Porath and T. Aulin, Algorithmic construction of trellis codes, *IEEE Trans. Commun.* COM-41(5): 649-654 (May 1993).
- J. H. Conway and N. J. A. Sloane, Sphere Packings, Lattices and Groups, Springer-Verlag, New York, 1988.
- G. D. Forney, Jr. et al., Efficient modulation for band-limited channels, *IEEE J. Select. Areas Commun.* SAC-2(5): 632–647 (1984).
- L. F. Wei, Trellis-coded modulation with multidimensional constellations, *IEEE Trans. Inform. Theory* **IT-33**: 483–501 (1987).
- A. R. Calderbank and N. J. A. Sloane, An eight-dimensional trellis code, *Proc. IEEE* 74: 757–759 (1986).
- L. F. Wei, Rotationally invariant trellis-coded modulations with multidimensional M-PSK, *IEEE J. Select. Areas Commun.* SAC-7(9): 1281–1295 (Dec. 1989).
- A. R. Calderbank and N. J. A. Sloane, New trellis codes based on lattices and cosets, *IEEE Trans. Inform. Theory* **IT-33**: 177–195 (1987).
- S. S. Pietrobon, G. Ungerböeck, L. C. Perez, and D. J. Costello, Jr., Rotationally invariant nonlinear trellis codes for two-dimensional modulation, *IEEE Trans. Inform. Theory* **IT-40**(6): 1773-1791 (Nov. 1994).
- S. S. Pietrobon et al., Trellis-coded multidimensional phase modulation, *IEEE Trans. Inform. Theory* **IT-36**: 63-89 (Jan. 1990).
- 43. IBM Europe, Trellis-coded modulation schemes for use in data modems transmitting 3-7 bits per modulation interval, CCITT SG XVII Contribution COM XVII, No. D114, April 1983.
- 44. L. F. Wei, Rotationally invariant convolutional channel coding with expanded signal space—Part I: 180 degrees, *IEEE J. Select. Areas Commun.* SAC-2: 659–672 (Sept. 1984).
- 45. L. F. Wei, Rotationally invariant convolutional channel coding with expanded signal space — Part II: Nonlinear codes, *IEEE J. Select. Areas Commun.* SAC-2: 672–686 (Sept. 1984).
- 46. IBM Europe, Trellis-coded modulation schemes with 8-state systematic encoder and 90° symmetry for use in data modems transmitting 3–7 bits per modulation interval, CCITT SG XVII Contribution COM XVII, No. D180, Oct. 1983.
- S. S. Pietrobon and D. J. Costello, Jr., Trellis coding with multidimensional QAM signal sets, *IEEE Trans. Inform. Theory* **IT-39**: 325–336 (March 1993).
- M. V. Eyuboglu, G. D. Forney, P. Dong, and G. Long, Advanced modem techniques for V. Fast, *Eur. Trans. Telecommun.* ETT-4(3): 234–256 (May–June 1993).
- G. D. Forney, Jr., Coset codes—Part I: Introduction and geometrical classification, *IEEE Trans. Inform. Theory* IT-34: 1123-1151 (1988).
- G. D. Forney, Jr., Coset codes Part II: Binary lattices and related codes, *IEEE Trans. Inform. Theory* **IT-34**: 1152–1187 (1988).
- G. D. Forney, Jr., Geometrically uniform codes, *IEEE Trans. Inform. Theory* IT-37: 1241–1260 (1991).
- D. Slepian, On neighbor distances and symmetry in group codes, *IEEE Trans. Inform. Theory* **IT-17**: 630-632 (Sept. 1971).
- 53. J. L. Massey, T. Mittelholzer, T. Riedel, and M. Vollenweider, Ring convolutional codes for phase modulation, *IEEE Int. Symp. Inform. Theory*, San Diego, CA, Jan. 1990.
- 54. P. Elias, Coding for noisy channels, *IRE Conv. Rec.* (Pt. 4): 37–47 (1955).
- R. Johannesson and Z.-X. Wan, A linear algebra approach to minimal convolutional encoders, *IEEE Trans. Inform. Theory* **IT-39**(4): 1219–1233 (July 1993).
- G. D. Forney, Jr., Convolutional codes I: Algebraic structure, IEEE Trans. Inform. Theory IT-16(6): 720–738 (Nov. 1970).
- J. E. Porath, Algorithms for converting convolutional codes from feedback to feedforward form and vice versa, *Electron. Lett.* 25(15): 1008–1009 (July 1989).
- J. P. Odenwalder, Optimal Decoding of Convolutional Codes, Ph.D. thesis, Univ. California, Los Angeles, 1970.
- K. J. Larsen, Short convolutional codes with maximum free distance for rates 1/2, 1/3, and 1/4, *IEEE Trans. Inform. Theory* **IT-19**: 371–372 (May 1973).
- E. Paaske, Short binary convolutional codes with maximum free distance for rates 2/3 and 3/4, *IEEE Trans. Inform. Theory* **IT-20**: 683-689 (Sept. 1974).

TRENDS IN BROADBAND COMMUNICATION NETWORKS

SASTRI KOTA Loral Skynet Palo Alto, California

1. INTRODUCTION

In the 1970s, the Advanced Research Projects Agency (ARPA) initiated the development of ARPANET, a very successful resource-sharing computer network [1,2]. ARPANET was a wide-area packet switching network, which later evolved into the Internet. This ARPANET resulted in a digital network resolution covering the telecommunication networks, data networks, and multimedia networks. The rapid growth of digital technologies allowed the integration of voice, data, and video to be processed, as a single stream and transported over global networks. The advances in high-speed networking broadband access technologies, the increasing power of the personal computer, the availability of information at the click of a button, and the exponential growth of the Internet, makes digital networks demand grow at a very rapid pace.

Traditionally, telecommunication networks utilized hierarchical controlled circuit switch technologies. Data networks demonstrated the strengths of packet-switched technology. The recent internetwork utilized various broadband services, such as voice, data, videostreams, multimedia, and group working, for schools, universities, hospitals, transported over either fiber, cable, satellite, and/or wavelength-division multiplexing (WDM). For example, the processor in a videogame is 10,000 times faster than Electronic Numerical Integrator and Computer (ENIAC) (1947) computer, and genesis game has more processing power than the 1976 Cray Super computer. Some of the chips used in some videocameras are more powerful than IBM 360. The increasing demand for broadband services, high-speed data transmission over the Internet and video on demand, require significant capacities challenging the fixed and mobile network designers and operations to deploy infrastructures to meet these future demands.

High speed Internet access was previously limited to enterprises, using technologies such as leased T-1, frame relay, or asynchronous transfer mode (ATM). But with the exponential growth of the Internet access for residential users, service providers have recognized the great opportunity in the residential broadband as well as more broadband enterprise markets.

It was reported that high-speed Internet access totaled 1.19 billion hours, 51% of the total 2.3 billion hours spent online, in January 2002. Twenty-two million used broadband to surf the Internet, an increase of 67% while enterprise usage increased by 42% [3]. As a result, the telephone, cable, and satellite companies have been developing xDSL, cable, and satellite access technologies. The objectives of these infrastructures are to provide higher bandwidth and speed to optimize the use of Internet for emerging applications such as content delivery distribution, e-finance, telemedicine, distance learning,

streaming video and audio, and interactive games. This article focuses mainly on the technology options for enterprise and residential users emphasizing the technical challenges through network examples.

This article is organized as follows. Broadband services, future applications and their requirements are discussed in Section 2. Section 3 provides a generic global communication model interconnecting the backbones based on either frame relay, ATM, or IP and even satellite technologies. The different broadband access technologies, their concepts and the enterprise access network solutions are described in Section 4. Section 5 discusses the residential broadband access technologies including Digital Subscriber Line (DSL), cable, wireless, and satellite solutions. Section 6 lists the standard organizations developing the protocols and interface standards, for enhancing the cost-effective interoperability of these infrastructures to meet the requirements of the growing high-speed applications. Section 7 describes the technical challenges for future networking with a multiprotocol label-switching service (MPLS)-based solution.

2. BROADBAND SERVICES AND APPLICATIONS

The Internet evolution has accompanied new opportunities for multimedia applications and services, ranging from simple file transfer to broadband access, IP multicast, media streaming, and content delivery distribution for both enterprise and consumer services. Figure 1 shows



Figure 1. Internet users. (*Source*: Nua Internet Surveys + vgc projections.)

Table 1. Broadband Services	Table 1.	Broadband	Services
-----------------------------	----------	-----------	----------

a projected growth of worldwide Internet users, where most users have demand for broadband services. In the consumer market the growing awareness of the Internet and activities ranging from shopping to finding local entertainment options to children's homework are driving the steep demand for more bandwidth. Education and entertainment content delivery has become one of the prime applications of Internet. During business globalization an increase of virtual business teams, enterprises, increase in competition for highly skilled workers, service providers and equipment vendors, are driving the demand for higher bandwidths or broadband.

Some of the observations are that 75% of traffic on the Internet is Web-based; there are 3.6 million Web sites with 300-700 million Web pages; the traffic consists of 80% data and 20% voice with a traffic growth of 100-1000% per year [4].

2.1. Broadband Services

Table 1 shows an example of the broadband services and applications. These include entertainment, broadband, and business services. A major challenge for these emerging services supported by Internet, which is an IP-based network, is to provide adequate QoS (quality of service). The normal QoS parameters from a user's perspective for both enterprise and residential users include throughput, packet loss, end-to-end delay, delay jitter, and reliability.

2.2. Network Requirements

The networking infrastructure supporting the enterprise and residential services should meet the following requirements:

• *High Data Rates*. The future applications such as videostreaming, media cast distributions, and high-speed Internet access for telemedicine applications and two-way telephonic education require rates ranging from a few hundred megabits to several gigabits. The three-generation (3G) systems cover up to 2 Mbps (megabits per second) for indoor environments and 144 kbps for vehicular environments. The IEEE 802.11-based broadband systems have approximately 20–30 Mbps transmission speeds. The data

Entertainment	Broadband	Business	Voice and Data Trunking
Broadcasting (DTH)	High-speed Internet access for consumer and enterprise	Telecommuting	IP Transport and ISP
Video on demand (VOD)	Electronic messaging	Videoconferencing	Voice over IP
Network or TV distribution	News on demand	E-finance, B2B	Video, audio, and data file transfer
TV cotransmissions	Multimedia	Home security	
Karaoke on demand	Distance learning	Unified messaging	
Games	MAN and WAN connectivities		
Gambling	Telemedicine		

rates for future generations will range from 2 to 600 Mbps depending on the system. The target speed for 4G cellular will be around 10-20 Mbps.

- *Delay*. Many real-time applications require minimum delay and the packet transfer delays for other classes of services are even stringent, reducing the queuing and processing delays.
- *Mobility*. The 4G cellular systems might be required to provide at least 2 Mbps for moving vehicles.
- *Wide Coverage*. The next-generation systems must provide good coverage area and roaming and handover to other systems.

Next-generation systems should provide at least an order of magnitude higher capacity, but the bit cost should be reduced to make the service more affordable. In addition to these requirements, the network must be scalable and provide security.

Figure 2 shows how different applications vary in their QoS requirements.

The most important QoS parameters are [5]

- *Throughput*. Throughput is the effective data transfer rate in bits per second (bps) of the network. Sharing of network capacity by a number of users reduces the throughput per user; as does the overhead of the packet. A service provider guarantees a minimum bandwidth rate.
- Delay. The time taken to transport data from the source to destination is known as *delay* or *latency*. For the public Internet, a voice call may easily exceed 150 ms of delay, because of processing delay and congestion. In GEO satellites one-way propagation delay can reach 250 ms. Such GEO satellite networks cannot service many real-time applications.
- *Jitter*. Jitter is caused by delay variation related to processing and queuing. These delay variations might be due to packet reassembly.
- *Reliability.* The percentage of network availability is an important parameter for the user. In satellitebased networks, the availability depends on the

Packet loss

frequency band of operation, power levels, antenna size, and the traffic for the service provided. Advanced error control techniques are used to provide good link availability [6].

Packet Loss. Packet loss occurs as a result of network congestion, error conditions, and link outages. Whenever buffers overflow, packets are dropped. Several buffer management schemes have been proposed to reduce the packet loss rate.

To meet future service requirements, a combination of old and new broadband networking technologies for enterprises include Gigabit Ethernet, frame relay, asynchronous transfer mode (ATM), multiprotocol label switching (MPLS), and broadband satellite networks. There are several options for residential users with respect to broadband access such as (1) dialup, (2) cable, (3) Digital Subscriber Loop (DSL) and its variants, (4) hybrid fiber-coax, (5) wireless including local multipoint distribution services (LMDS) and multichannel multipoint distribution services (MMDS), (6) satellite access, and (7) leased lines.

3. GLOBAL NETWORK MODEL

The tremendous success and exponential growth of the Internet and the new applications such as broadcasting, multicasting, and video distribution have resulted in significantly higher data rates. One of the key requirements for the emerging "global network," which is a "network of networks," is rich connectivity among fixed as well as mobile users. Advances in switching and transport technologies have made increase in transmission bandwidth and switching speeds possible, and still more dramatic increases are possible via optical switching [7]. The future generation of communications networks provides "multimedia services," "wireless (cellular and satellite) access to broadband networks," and "seamless roaming among different systems."

Figure 3 shows a global communication network scenario providing connectivity among corporate networks, Internet, and the ISPs. The networking technologies vary



Figure 2. Application-specific QoS requirements example.



Figure 3. Communication network scenario.

can vary between ATM, frame relay, IP and optical backbones. The access technologies could be dialup, cable, DSL, and satellite.

Mobile communications are supported by secondgeneration digital cellular (GSM); data service is supported by GPRS. Third-generation systems such as IMT-2000 can provide 2 Mbps and 144 kbps indoors and in vehicular environments. Even 4G and 5G systems are being studied to provide data rates 2-20 Mbps and 20-100 Mbps, respectively. See Section 5.4 for further details.

Several broadband satellite networks at Ka band are planned and being developed to provide such global connectivity for both fixed satellite service (FSS) and mobile satellite service (MSS) using geosynchronous (GSO) and nongeosynchronous (NGSO) satellites as discussed in Section 4.5. Currently GSO satellite networks with verysmall-aperture terminals (VSATs) at Ku bands are being used for several credit card verifications, rental cars, banking, and other applications. Satellite networks such as StarBand, Direway, and WildBlue are being developed for high speed Internet access (see Section 5.3).

4. BROADBAND ENTERPRISE NETWORKING

This section describes the networking technologies for enterprise including the use of Gigabit Ethernet, frame relay, ATM, IP, and broadband satellite. The technology concepts and advantages are discussed with some examples.

4.1. Gigabit Ethernet

Gigabit Ethernet is an extension of the IEEE 802.3 Ethernet standard. It builds on the Ethernet protocol

but increases speed 10-fold over Fast Ethernet to 10 Gbps. In March 1999, a working group was formed at IEEE 802.3 to develop a standard for 10-gigabit Ethernet. Gigabit Ethernet is basically the fasterspeed version of Ethernet. It will support the data rate of 10 Gbps and offers benefits similar to those of the preceding Ethernet standard [8]. The potential applications for 10-gigabit Ethernet enterprise users are universities, telecommunication carriers, and Internet service providers. One of the main benefits of the 10gigabit standard is that it offers a low-cost solution to solve the demand on bandwidth. In addition to the low cost of installation, the cost of network maintenance and management is minimal. Local network administrators manage and maintain for 10-gigabit Ethernet networks.

In addition to the cost reduction benefit, 10-gigabit Ethernet allows for faster switching and scalability. 10gigabit Ethernet uses the same Ethernet format, it allows seamless integration of LAN, MAN, and WAN. There is no need for packet fragmentation, reassembling, or address translation, eliminating the need for routers that are much slower than switches. 10-gigabit Ethernet also offers straightforward scalability since the upgrade paths are similar to those of 1-gigabit Ethernet.

In LAN markets, applications typically include inbuilding computer servers, building-to-building clusters, and data centers. In this case, the distance requirement is relaxed, usually between 100 and 300 m. In the mediumhaul market, applications usually include campus backbones, enterprise backbones, and storage area networks. In this case, the distance requirement is moderate, usually between 2 and 20 kms. The cabling infrastructure usually already exists. The technologies must operate over it. The initial cost is not so much of an issue. Normally, users are willing to pay for the cost of installation but not the cost of network maintenance and management. Ease of technology is preferred. This gives an edge to 10-gigabit Ethernet over the only currently available 10-gigabit SONET technology (OC-192c).

WAN markets typically include Internet service providers and Internet backbone facilities. Most of the access points for long-distance transport networks require the OC-192c data rate. So the key requirement for these markets is the compatibility with the existing OC-192c technologies. Mainly, the data rate should be 9.584640 Gbps. Thus, the 10-Gigabit Ethernet standard specifies a mechanism to accommodate the rate difference.

4.1.1. Protocol. The Ethernet protocol basically implements the bottom two layers of the Open Systems Interconnection (OSI) seven-layer model, that is, the data-link and physical sublayers. Figure 4 depicts the typical Ethernet protocol stack and its relationship to the OSI model.



Figure 4. Ethernet protocol layer.

- Media Access Control (MAC). The media access control sublayer provides a logical connection between the MAC clients of itself and its peer station. Its main responsibility is to initialize, control, and manage the connection with the peer station.
- *Reconciliation Sublayer.* The reconciliation sublayer acts as a command translator. It maps the terminology and commands used in the MAC layer into electrical formats appropriate for the physicallayer entities.
- 10-Gigabit Media-Independent Interface (10GMII). 10GMII provides a standard interface between the MAC layer and the physical layer. It isolates the MAC layer and the physical layer, enabling the MAC layer to be used with various implementations of the physical layer.
- *Physical Coding Sublayer (PCS).* The PCS sublayer is responsible for coding and encoding data streams to and from the MAC layer. A default coding technique has not been defined.
- *Physical Medium Attachment (PMA).* The PMA sublayer is responsible for serializing code groups into bit streams suitable for serial bit-oriented physical devices and vice versa. Synchronization is also done for proper data decoding in the sublayer.
- *Physical Medium-Dependent (PMD).* The PMD sublayer is responsible for signal transmission. The typical PMD functionality includes amplifier, modulation, and waveshaping. Different PMD devices may support different media.
- *Medium-Dependent Interface (MDI).* MDI is referred to as a connector. It defines different connector types for different physical media and PMD services.

4.1.2. Fiber Network Architecture. Figure 5 shows Gigabit Ethernet used in an evolving fiber enterprise



Figure 5. Gigabit Ethernet in a fiber enterprise network.

network configuration. The service provider backbone consists of an OC-48 (2.48 Gbps) or an OC-192 (10 Gbps) using dense wavelength-division multiplexing (DWDM) with connections to ATM over SONET and packet over SONET networks via core routers. The access networks could use OC-3 (155.52 Mbps), or T3 (44.736 Mbps), or DS1 (1.544 Mbps). A typical application using such an enterprise network solution could be distribution of analog and digital video.

4.2. Frame Relay

Frame relay is a standard communication protocol that is specified in ITU-T (formerly CCITT) recommendations I.122 and Q.922, which add relay and routing functions to the data-link layer of the OSI model [9]. Subsequently, the Frame Relay Forum has developed the frame relay specification for wide-area networks. Frame relay services were developed by service providers for enterprise as a cost effective and a better flexible alternative to time-division multiplexing (TDM) and private line services. Enterprises needed dedicated connectivity between offices but could not necessarily afford dedicated circuits. Meanwhile service providers required a reliable means to subscribe their bandwidth-constrained networks.

The frame relay protocol has been particularly effective for data traffic. Carriers generally use frame relay as access technology and ATM as a transport. The network architecture requires carriers to maintain a completely dedicated ATM/frame relay network in addition to their IP and voice networks.

The rapid increase in high bandwidth communication is the main reason for using frame relay technology. Two main factors influence the rapid demand for highspeed networking: (1) rapid increase in use of LANs, and (2) use of fiber optic links. Frame relay is a packetswitching technology, which relies on low error rate digital transmission links and high-performance processors. Frame relay technology was designed to cover (1) low latency and higher throughput, (2) bandwidth on demand, (3) dynamic sharing of bandwidth, and (4) backbone network.

For enterprises, frame relay is a well-understood technology, and by definition, is a layer 2 technology that supports oversubscription. The frame relay technology has some design advantages as well as restrictions including

- Unpredictable bandwidth and maximum speed capacity at DS3
- Hierarchical aggregation schemes that use hub and spoke architectures
- Scaling complexities by having to add additional layer 2 addresses to different sites rather than by IP's inherent self-healing and learning capabilities
- Used for interconnecting LANs and particularly WANs, and more recently, for voice and videocon-ferencing
- Provides LAN-to-LAN connectivity from 56 kbps to 1.5 Mbps
- Offers congestion control and higher performance

4.2.1. Frame Relay Data Unit. The frame structure in a frame relay network consists of two flags indicating the beginning and the end of the frame, an address field, an information field, and a frame check sequence [10]. In addition to the address, the address field contains functions that warn of overload and indicate which frames should be discarded first. The fields and their purpose are discussed in detail below.

- *Flag.* All frames begin and end with a flag consisting of an octet composed of a known bit pattern: a zero followed by 6 ones and a zero (01111110).
- *Address*. In the two octets in the address field, the first 6 bits of the first octet and the first 4 bits of the second octet are used for addressing. These 10 bits, which form the DLCI, select the next destination to which the frame is to be transported.
- *CR*. Command response is not used by the frame relay protocol. It is sent transparently through the frame relay nodes and can be made available to users as required.
- *EA*. At the end of each address octet there is an extended address bit that can allow extension of the DLCI field to more than 10 bits. If the EA bit is set to "0," another address octet will follow. If it is set to "1," the octet in question is the last one in the address field.
- *FECN*. If overload occurs in the network, forward explicit congestion notification is indicated to alert the receiving end. The network makes this indication, and end users need not take any specific action.
- *BECN*. Similar to FECN, backward explicit congestion notification alerts the sending end to an overload situation in the network.
- *DE*. Discard eligibility indicates that the frame is to be discarded in case of overload. This indication can be regarded as a prioritizing function, although frames without a DE indication can also be discarded.
- *Information Field.* This is where we find user information. The network operator decides how many octets the field is allowed to contain, but the Frame Relay Forum recommends a maximum of 1600. The information passes through the network completely unchanged and is not interpreted by the frame relay protocol.
- *FCS*. The frame check sequence checks the frame for errors. All bits in the frame, except the flags and FCS, are checked.

The frame relay frame format is shown in Fig. 6.

4.2.2. Frame Relay Examples

4.2.2.1. Frame Relay LAN-to-LAN. Figure 7 shows LAN-to-LAN and an ISP connectivity using frame relay at 56 kbps, 384 kbps, T1, and T1/T3, respectively.

4.2.2.2. Frame Relay WAN Architecture. Figure 8 shows a frame relay network with various offices connected via frame relay to four aggregation hubs. Depending on







Figure 7. Frame relay overview.

Figure 8. WAN architecture.

the size of the organization and the speeds at which different regional offices connect, the hub will have at least two large WAN routers. For some organizations, hubs may be fed by hundreds of regional locations through frame relay or private line connections. Although some of the traffic may terminate at its locally connected "hub" data center, most of the traffic "hairpins" in and out of the local data center on its way to a remote hub or destination. In this case, the hub sites provide statistical multiplex gains.

4.3. Asynchronous Transfer Mode (ATM)

ATM is an International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) standard for cell relay wherein information for multiple service types, such as voice, video, or data, is conveyed in small, fixed-size cells. ATM networks are connection oriented. ATM is based on the efforts of the ITU-T Broadband Integrated Services Digital Network (BISDN) standard. It was originally conceived as a high-speed transfer technology for voice, video, and data over public networks. The ATM Forum extended the ITU-T's vision of ATM for use over public and private networks.

ATM is a cell switching and multiplexing technology that combines the benefits of circuit switching (guaranteed capacity and constant transmission delay) with those of packet switching (flexibility and efficiency for intermittent traffic) [11]. It provides scalable bandwidth from a few megabits per second (Mbps) to many gigabits per second (Gbps). Because of its asynchronous nature, ATM is more efficient than synchronous technologies, such as time-division multiplexing (TDM). With TDM, each user is assigned to a time slot, and no other station can send in that time slot. If a station has a lot of data to send, it can send only when its time slot comes up, even if all other time slots are empty. If, however, a station has nothing to transmit when its time slot comes up, the time slot is sent empty and is wasted. Because ATM is asynchronous, time slots are available on demand with information identifying the source of the transmission contained in the header of each ATM cell.

4.3.1. ATM Reference Model. Figure 9 illustrates the B-ISDN Protocol Reference Model, which is the basis for the protocols that operate across the User Network Interface (UNI). The B-ISDN reference model consists of three planes: the user plane, the control plane, and the management plane. Reference ITU-T I.121 describes the ATM reference model and various functions in detail.

4.3.2. ATM Cell Format. An ATM cell consists of 48 bytes of data with a 5-byte header as shown in Fig. 10 [11]. The cell size was determined by ITU-T as



Figure 9. ATM protocol architecture.

a compromise between voice and data requirements. The header fields are as follows:

- *Generic Flow Control* (*GFC*). Provides local functions, such as identifying multiple stations that share a single ATM interface.
- *Virtual Path Identifier (VPI)*. In conjunction with the VCI, identifies the next destination of a cell as it passes through a series of ATM switches on the way to its destination.
- *Virtual Channel Identifier (VCI)*. In conjunction with the VPI, identifies the next destination of a cell as it passes through a series of ATM switches on the way to its destination.
- *Payload Type (PT)*. If the cell contains user data, the second bit indicates congestion, and the third bit indicates whether the cell is the last in a series of cells that represent a single AAL5 frame.
- *Cell Loss Priority* (*CLP*). Indicates whether the cell should be discarded if it encounters extreme congestion as it moves through the network. If the CLP bit equals 1, the cell should be discarded in preference to cells with the CLP bit equal to zero.
- *Header Error Control (HEC)*. Calculates checksum only on the header.

There are two types of ATM services: permanent virtual circuits (PVCs) and switched virtual circuits (SVCs). A PVC allows direct static connectivity between sites similar to a leased line. It guarantees availability of a connection and does not require a signaling protocol. On the other hand, SVC allows dynamic setup and release of connections. Dynamic call control requires a signaling protocol between the ATM endpoint and the switch. This service provides flexibility. However, it results in a signaling overhead in setting up the connection.

4.3.3. Classes of Service. Table 2 provides different classes of network traffic that need to be treated differently by an ATM network [12].

4.3.4. Traffic Management and QoS. One of the significant advantages of ATM technology is providing QoS guarantees as described in the ATM Forum's *Traffic Management Specification* [13]. The framework supports five service categories, namely, constant bit rate (CBR), real-time variable bit rate (rt-VBR), non-real-time VBR (nrt-VBR), unspecified bit rate (UBR), and available bit



Figure 10. ATM cell structure.

	Constant-Bit-Rate (CBR) Service	Real-Time Service (VBR-rt)	Non-Real-Time Connection-Oriented Data Service (VBR-nrt-COD)	Non-Real-Time Connectionless Data Service (VBR-nrt-CLS)	Unspecified Bit Rate (UBR)	Available Bit Rate (ABR)
Bearer class	Class A	Class B	Class C	Class D	Class X	Class Y
Applications	Voice and clear channel	Packet video and voice	Data	Data	Data	Data
Connection mode	Connection oriented PVC or SVC	Connection oriented PVC or SVC	Connection oriented PVC or SVC	Connection less	Connection oriented	Connection oriented
Bit rate	Constant	Variable	Variable	Variable	Variable	Variable
Timing required	Required	Required	Not required	Not required	Not required	Not required
Services	Private line	None	Frame relay	SMDS	Raw cell	
AAL	1	2	3/4 & 5	3/4	Any	3/4 & 5

Table 2. Classes of Service

rate (ABR), with an additional one, guaranteed frame rate (GFR). With the exception of UBR, all ATM service categories require incoming traffic regulation to control network congestion and ensure QoS guarantees. This function is performed by access policing devices to determine whether the traffic conforms to certain traffic characterizations. The conformant cells are allowed to enter the ATM network and receive QoS guarantees, whereas the nonconformant cells will be either dropped or tagged. Tagged cells may be allowed into the network but will not receive any QoS guarantees. The other traffic management functions include connection admission control (CAC), traffic shaping, usage parameter control (UPC), resource management, priority control, cell discarding, and feedback controls. The ATM Forum [13] provides the details of these functions and traffic management algorithm recommendations for different service classes. Table 3 provides the various traffic parameters and the QoS parameters.

4.4. IP Enterprise Network

Since the late 1990s the Internet has become the major vehicle for most of the broadband applications and for the growth of the telecommunication network. The Internet protocol has become the universal network layer protocol

Table 3. ATM Service Category Attributes

for both wireline and wireless networks. At the transport layer, Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) have become the most popular ones. The UDP has been attracted with the multimedia services. In this section, an overview of TCP and IP protocols with formats and examples is described.

4.4.1. TCP/IP Protocol. The TCP/IP protocol suite is the mostly used Internet protocol for global networks. Many of the Internet applications such as file transfer, email, Web browsing, streaming media, and newsgroups use TCP/IP protocols. Figure 11 shows the TCP/IP protocol stack with respect to ISO/OSI protocol. The overhead introduced at each layer to the application datagram is also described.

4.4.1.1. Transmission Control Protocol (TCP). The TCP provides reliable transmission of data in an IP environment. TCP corresponds to the transport layer (layer 4) of the OSI reference model [14]. Among the services TCP provides are stream data transfer, reliability, efficient flow control, full-duplex operation, and multiplexing. TCP offers reliability by providing connection-oriented, end-toend reliable packet delivery through an internetwork by using acknowledgments. The reliability mechanism of TCP

	Traffic Parameters			QoS	QoS Parameters		
Service Categories	PCR, CDVT _{PCR}	$\begin{array}{c} {\rm SCR,\ MBS,}\\ {\rm CDVT}_{\rm SCR} \end{array}$	MCR	Peak-to-Peak CDV	Maximum CTD	CLR	Others
CBR	Yes	No	No	Yes	Yes	Yes	No
rt-VBR	Yes	Yes	No	Yes	Yes	Yes	No
nrt-VBR	Yes	Yes	No	No	No	Yes	No
UBR	Yes	No	No	No	No	No	No
ABR	Yes	No	Yes	No	No	No	Feedback
GFR	Yes	MFS, MBS	Yes	No	No	No	No

Key: PCR—peak cell rate; SCR—sustainable cell rate; MCR—minimum cell rate; MBS—maximum burst size; CDVT—cell delay variation tolerance; CDV—cell delay variation; CTD—cell transfer delay; CLR—cell loss ratio; MFS—maximum frame size.

2662 TRENDS IN BROADBAND COMMUNICATION NETWORKS

OSI model	TCP/IP
Application	Application (SMTP, NNTP,
Presentation	FTP, SNMP, HTTP)
Session	
	Transport
Transport	(TCP, UDP)
Network	Network (IP, ICMP, IGMP)
Data link	
	Physical
Physical	

TCP - Transmission Control Protocol

UDP – User Datagram Protocol

ICMP - Internet Control Message Protocol

IGMP - Internet Group Management Protocol

IP – Internet Protocol

SMTP – Simple Mail Transfer Protocol

NNTP – Network News Transfer Protocol

FTP – File Transfer Protocol

SNMP – Simple Network Management Protocol

HTTP – Hyper Text Transfer Protocol



AH – Application Header TCPH – TCP Header (20 bytes) IPH – IP Header (20 bytes) EH – Ethernet Header

Figure 11. TCP/IP protocol stack.

Source port (16)			Destination	port (16)
Sequence			umber (32)	
Acknowledgement number (32)				
HeaderReserved6 Controlsize (4)(6)bits (6)			Window s	ize (16)
Checksum (16)			Urgent poir	nter (16)
Options				Padding
Data (variable length)				

Figure 12. TCP segment format.

Source port (16)	Destination port (16)
Length (16)	Checksum (16)
Data (varia	ble length)

Figure 13. UDP segment format.

allows devices to deal with lost, delayed, duplicate, or misread packets. The lost packets are detected by a timeout mechanism. TCP offers a efficient flow control mechanism by using sliding windows to avoid buffer overflow. TCP provides a full-duplex operation and multiplexing of data. Along with the IP Security protocol (IPSec), Internet security can be obtained. Figure 12 shows the format of a TCP segment. The header is 20 bytes long.

4.4.1.2. User Datagram Protocol (UDP). UDP is a connectionless transport-layer protocol that enables best-effort datagrams to be transmitted between host system

applications. Unlike TCP, UDP adds no reliability, flow control, or error recovery functions to IP [14]. It uses an 8-byte header. The format of a UDP segment is shown in Fig. 13.

4.4.1.3. Internet Protocol (IP). IP is a connectionless protocol working at the network layer (layer 3). IP has two primary responsibilities: providing connectionless "best effort" delivery of datagrams across and between networks; and providing fragmentation and reassembly of datagrams to support data links with different maximum transmission unit (MTU) sizes.

TRENDS IN BROADBAND COMMUNICATION NETWORKS 2663

Version (4)	Header size (4)	Type of service (8)	Total length (16)	
Identification (16)		Flags (3) Fragmentation offset (13)		
Time to	live (8)	Protocol (8)	Header checksum (16)	
	Source IP address (32)			
		Destination IF	9 address (32)
Options				
Data (variable length)				

Figure 14. IPv4 segment format.

Version (4)	Traffic class (8)	Flow label (20)			
Payload length (16)			Next header (8)	Hop limit (8)	
	Source address (128)				
Destination address (128)					
Data (variable length)					

Figure 15. IPv6 segment format.

4.4.2. IPv4 and IPv6 Formats. Currently most of the Internet uses IP version 4 (IPv4) [14]. IP, as originally developed, has no reliability or QoS mechanisms. A new version, IP version 6 (IPv6) has been designed as the successor of IPv4. IPv6 provides more QoS and address space than IPv4 [15]. Figures 14 and 15 show the formats of IPv4 and IPv6 segments. The changes from IPv4 to IPv6 fall into the following categories:

- *Expanded Address Capabilities*. IPv6 increases the IP address size from 32 to 128 bits, to support more levels of addressing hierarchy, and a much greater number of addressable nodes.
- *Header Format Simplification*. Some IPv4 header fields have been dropped or made optional, to reduce the processing cost of packet handling.
- Improved Support for Extensions and Options. Changes in the way IP header options are encoded allows for more efficient forwarding and greater flexibility for introducing new options in the future.
- *Flow Labeling Capability*. A new capability is added to enable the labeling of packets belonging to a particular traffic "flow" for which the sender requests special handling.

The fields in the IPv6 segment are

- *Version*—4-bit Internet Protocol version number = 6.
- *Traffic class* 8-bit traffic class field. Available for use by originating nodes and/or forwarding routers

to identify and distinguish between different classes or priorities of IPv6 packets.

- *Flow label*—20-bit flow label. Used by a source to label sequences of packets for which it requests special handling by the IPv6 routers.
- *Payload length* 16-bit unsigned integer. Length of the rest of the packet following the header.
- *Next header*—8-bit selector. Identifies the type of header immediately following the IPv6 header.
- *Hop limit*—8-bit unsigned integer. Decremented by 1 by each node that forwards the packet. The packet is discarded if hop limit is decremented to zero.
- Source address 128-bit address of the originator of the packet.
- *Destination address* 128-bit address of the intended recipient of the packet.

4.4.3. Network Example. As a less expensive WAN technology than frame relay, IP virtual private networks (IPVPNs) are being developed. The premise behind the IPVPNs is to provide a logically private network across a shared infrastructure using tunneling or IPSec protocol to provide security to the enterprise network [16]. Figure 16 shows such an IPVPN network for an enterprise solution consisting of a number of virtual routers from the different sites connected together across the Internet. These virtual routers are used to enable the VPN services connecting all sites in a mesh topology.

The advantage of IPVPNs is that the public Internet provides ubiquitous connectivity. IPSec is used to tunnel the data between the sites across an entirely besteffort public network. The corporations use IPVPNs



Figure 16. IP enterprise network model.

to connect their data centers, remote offices, mobile employees, telecommuters, customers, suppliers, and business partners. The IPVPN solutions is very attractive because it offers a dedicated private network and is relatively cost-effective.

Many large enterprises have not deployed IPVPNs as expected for the following reasons:

- *Virtual Routers*. The IPVPNs use a virtual router per VPN and the use of virtual router has disadvantages such as (1) no autonomy due to the use of public backbone, (2) enterprise does not have any control on the traffic flows across the network, and (3) the performance of the virtual router is inversely proportional to the number of routers in the network.
- Service-Level Agreement. It is very difficult for an enterprise to have a reasonable SLA.
- Security. In the enterprise solution, security protocols like IPSec are employed to secure the data transported over a public Internet. However, large corporations normally spend a good portion of their resources monitoring and improving security of their network.
- *Service Guarantees*. There is no possibility of bandwidth or connectivity guarantee when the traffic is carried over a public Internet comprised of many service providers.

Although the IPVPN solution for enterprise network discussed here provides cheap connectivity, it is not well adopted by large enterprises because of the nature of the use of public backbone. It is not suitable for high-bandwidth applications. IPVPNs have been adopted, however, as a cheap solution for network connectivity for small-medium-sized companies.

4.5. Broadband Satellite Networks

Because of the inherent broadcast and multiple access characteristics, satellites have become an attractive solution for enterprise and consumer users. This section describes the future satellite systems operating at the C, Ku, and Ka bands for enterprise applications [17].

4.5.1. Satellite Network Characteristics. The principal attributes of satellites include (1) broadcasting and (2) long-haul communications. For example, Ka-band spot beams generally cover a radius of 200 km (125 mi) or more, while C- and Ku-band beams can cover entire continents. This is in contrast to last-mile-only technologies that generally range from 2 to 50 km (or 1-30 mi).

The main advantages of satellite communications are [18]

- Ubiquitous Coverage. A single satellite system can reach every potential user across an entire continent regardless of location, particularly in areas with low subscriber density and/or otherwise impossible or difficult to reach. Current satellites have various antenna types that generate different footprint sizes. The sizes range form coverage of the whole earth as viewed from space (about a third of the surface) down to a spot beam that covers much of Europe or North America. All these coverage options are usually available on the same satellite. Selection between coverage is made on transparent satellites by the signal frequencies. It is spot beam coverage that is most relevant for access since they operate to terminal equipment of least size and cost. Future systems will have very narrow spot beams of a few hundred miles across that have a width of a fraction of a degree.
- *Bandwidth Flexibility*. Satellite bandwidth can be configured easily to provide capacity to customers in virtually any combination or configuration required. This includes simplex and duplex circuits from narrowband to wideband and symmetric and asymmetric configurations. Future satellite networks with narrow spot beams are expected to deliver rates of up to 100 Mbps with 90-cm antennas, and the backplane speed within the satellite switch could be typically in the Gbps range. The uplink rate from a 90-cm user terminal is typically 384 kbps.

- *Cost*. The cost is independent of distance; the wide area coverage from a satellite means that it costs the same to receive the signal from anywhere within the coverage area.
- *Deployment*. Satellites can initiate service to an entire continent immediately after deployment, with short installation times for customer premise equipment. Once the network is in place, more users can be added easily.
- *Reliability and Security*. Satellites are amongst the most reliable of all communication technologies, with the exception of SONET fault-tolerant designs. Satellite links require only that the end stations be maintained, and they are less prone to disabling though accidental or malicious damage.
- *Disaster Recovery*. Satellite communication provides an alternative to damaged fiberoptic networks for disaster recovery options and provide emergency communications.

4.5.2. Market Potential. Figure 17 shows the market potential for satellite-based broadband enterprise and residential access users. The enterprise market is expected to grow up to \$4 billion by 2006 for content delivery distribution (CDD) [19].

4.5.3. Next-Generation Ka Band. Until the late 1990s, the Ka band was used for experimental satellite programs in the United States, Japan, Italy, and Germany. In the United States, the Advanced Communications Technology Satellite (ACTS) is being used to demonstrate advanced technologies such as onboard processing and scanning spot beams. A number of applications were tested, including distance learning, telemedicine, credit card financial transactions, high-data-rate computer interconnections, videoconferencing, and high-definition television (HDTV). The growing congestion of the C and Ku bands and the success of the ACTS program increased the interest of satellite system developers in the Kaband satellite communications network for exponentially growing Internet access applications. A rapid convergence of technical, regulatory, and business factors has increased the interest of system developers in Ka-band frequencies.

Several factors influenced the development of multimedia satellite networks at Ka-band frequencies:

- Adaptive Power Control and Adaptive Coding. Adaptive power control and adaptive coding technologies have been developed for improved performance, mitigating propagation error impacts on system performance at the Ka band.
- *High Data Rate.* A large bandwidth allocation to geosynchronous fixed satellite services (GSO FSS) and nongeosynchronous fixed satellite services (NGSO FSS) makes high-data-rate services feasible over Ka-band systems.
- Advanced Technology. Development of low-noise transistors operating in the 20-GHz band and high-power transistors operating in the 30-GHz band have influenced the development of low-cost earth terminals. Space-qualified higher-efficiency traveling-wave tubes (TWTAs) and ASICs development have improved the processing power. Improved satellite bus designs with efficient solar arrays and higher efficiency electric propulsion methods resulted in cost-effective launch vehicles.
- *Global Connectivity*. Advanced network protocols and interfaces are being developed for seamless connectivity with terrestrial infrastructure.
- *Efficient Routing*. Onboard processing and fast packet or cell switching (e.g., ATM, IP) makes multimedia services possible.
- *Resource Allocation*. Demand assignment multiple access (DAMA) algorithms along with traffic management schemes provide capacity allocation on a demand basis.
- *Small Terminals*. Multimedia systems will use small and high-gain antenna on the ground and on the satellites to overcome path loss and gain fades.
- *Broadband Applications*. Ka-band systems, combining traditional satellite strengths of geographic reach and high bandwidth, provide the operators a large subscriber base with scale of economics to develop consumer products.

Figure 18 illustrates a broadband satellite network architecture represented by a ground segment, a space segment, and a network control segment. The ground segment consists of terminals and gateways (GWs),



Figure 17. Satellite CDD service growth. (Source: Pioneer Consulting 2002.)



Figure 18. Broadband satellite network configuration.

NCS - Network Control Station UNI - User Network Interface

which may be further, connected to other legacy public and/or private networks. The network control station (NCS) performs various management and resource allocation functions for the satellite media. Intersatellite crosslinks in the space segment to provide seamless global connectivity via the satellite constellation are optional. Hybrid network architecture allows the transmission of packets over satellite, and multiplexes and demultiplexes datagram streams for uplinks, downlinks, and interfaces to interconnect terrestrial networks. The satellite network configuration also illustrates the signaling protocol (e.g., SS7, UNI) and the satellite interface unit. The architectural options could vary from ATM switching, IP transport, to MPLS over satellite.

4.5.4. Future Broadband Satellite Networks. The satellite network architectural options are

- GSO versus NGSO
- No onboard processing or switching
- Onboard processing with ground-based cell or ATM switching or fast packet switching
- Onboard processing and onboard ATM or fast packet switching

The first-generation services that are now in place use existing Ku-band fixed satellite service (FSS) for two-way connections. Using FSS a large geographic area is covered by a single broadcast beam. The new Ka-band systems use spot beams that cover a much smaller area, say, hundreds of miles across. Adjacent cells can use a different frequency range, but a given frequency range can be reused many times over a wide geographic area. The frequency reuse in the spot beam technology increases the capacity. In general, Ka spot beams can provide 30–60 times the system capacity of the first-generation networks.

The next-generation satellite multimedia networks can be divided into two classes: (1) the broadband satellite *connectivity network*, in which full end-toend user connectivity was established—the proposed global satellite connectivity networks such as Astrolink, Spaceway, and EuroSkyway have onboard processing and switching capabilities; and (2) regional access networks, such as StarBand, IPStar, and WildBlue, which are intended to provide Internet access. These access systems employ nonregenerative payloads.

In a nonregenerative architecture, the satellite receives the uplink and retransmits it on the downlink without onboard demodulation or processing. In a processing architecture with cell switching or layer 3 package, the satellite receives the uplink, demodulates, decodes, switches, and buffers the data to the appropriate beam after encoding and remodulating the data, on the downlink. In a processing architecture, switching and buffering are performed on the satellite; in a nonprocessing architecture, switching/routing and buffering are performed within a gateway. A nonregenerative architecture has physicallayer flexibility but limited, if any, network-layer flexibility. A process architecture has limited physical-layer flexibility, but greatly increased network-layer flexibility, and permits any network topology from point-to-point, hub-and-spoke architecture. The selection of the satellite network architecture is strictly dependent on the target customer applications and performance/cost tradeoffs.

Table 4 compares some of the new-generation C-, Ka-, and Ku-band satellite systems. These systems, which are under development, provide global coverage and high bandwidth.

5. RESIDENTIAL BROADBAND ACCESS

This section describes residential access technologies including DSL, hybrid fiber-coax, fixed wireless, satellite access, and mobile wireless. The technology and examples are described [20].

5.1. Residential Access Market Potential

The different access technologies for broadband applications include cable modem, xDSL, satellite, wireless, and fiber to the home. A report from ARC group concluded that the residential broadband market would be worth \$80 billion by 2007 with nearly 300 million residential and

Services	Spaceway	$\operatorname{Astrolink}^a$	EuroSky Way	Teledesic	Intelsat	Eutelsat
Data uplink	384 kbps-6 Mbps	384 kbps-2 Mbps	160 kbps-2 Mbps	16 kbps-2 Mbps	_	$\leq 2 \text{ Mbps}$
Data downlink	384 kbps-20 Mbps	384 kbps - 155 Mbps	128-640 kbps	16 kbps-64 Mbps	$\leq\!45~Mbps$	$55 \mathrm{~Mbps}$
Number of satellites	8	9 (4 initially)	5	30	—	—
Satellite	GEO	GEO	GEO	MEO	GEO	GEO (Hotbird 3-6)
Frequency band	Ka	Ka	Ka	Ka	C, Ku	Ku, Ka
Onboard processing	Yes	Yes	Yes	Yes	No	_
Operation scheduled	2003	2003	2004	2004/5	_	2001

 Table 4. Global Broadband Satellite Networks

^aProgram currently on hold.

commercial sites using broadband. Nearly a third of broadband connections will be DSL, with cable close behind. Satellite, wireless, and others will fill the remainder. In fact, for broadband satellite, the high growth rate for CDD applications by 2006 is projected to be \$1.45 billion and the revenue growth for wireless, about \$3.3 billion in for that same year. It is estimated that 3.8 million subscribers for DSL in 2001 will grow to 20.7 million subscribers in 2006, and to 21 million subscribers to cable modem, according to a Pioneer Consulting report [19].

5.2. Broadband Access Technologies

Figure 19 shows the available broadband access technology options. The Asymmetric DSL (ADSL) provides upstream at 64 kbps, and downstream at 1.8 Mbps whereas very high bit rate (VDSL) supports 26 Mbps symmetric, and in the asymmetric VDSL supports less than 6.4 Mbps upstream and less than 52 Mbps downstream [20]. A digital subscriber line access multiplexer (DSLAM) delivers high-speed data transmission over the existing copper telephone lines. The DSLAM separates the voice frequency signals from high-speed data traffic and controls and routes xDSL traffic between the subscriber's end-user equipment (router, modem, or network interface card) and the network service provider. The cable head end, which includes cable modem termination system, enables high-speed Internet access. Details of the cable modem protocol and access are described in Section 5.2.2.

Table 5 provides a set of advantages and disadvantages of the various broadband access technologies. The technologies compared are satellite, hybrid fiber-coax, DSL, and LMDS/MMDS. The different access technologies are described in the following text.

5.2.1. Digital Subscriber Line (DSL) Access. Digital subscriber line (DSL) is a technology that uses regular telephone lines to transmit a high volume of data at a very high speed. The telephone uses only part of the frequency available on these copper lines; DSL gets more from them by splitting the line—using the higher frequencies for data, the lower for voice and fax [21–26].

DSL offers high-speed broadband connectivity over existing copper telephony infrastructure. Upgrading copper, LECs can offer telephony and data services simultaneously.

5.2.1.1. Key Benefits of DSL

• Compared to a 56K modem, DSL speeds range from twice as fast up to 125 times as fast.



Figure 19. Broadband access options for residential service.

2668 TRENDS IN BROADBAND COMMUNICATION NETWORKS

Access Technology	Advantages	Disadvantages
Satellite	No router hops	High installation cost
	Excellent for content distribution and management through broadcast and multicast	Adequate TCP/IP performance
	Excellent video quality	Signal propagation delay limits real-time applications
	Experiences less packet loss than terrestrial options	
Hybrid fiber–coax	Established presence in most residences — no need to draw new cables	Low security
	Good picture quality	Shared bandwidth may limit performance
	Good customer satisfaction for service	Limited enterprise market coverage
		Prone to external noise degradation
DSL	Reuse of existing copper	Access rate is function of distance from central office
	Dedicated bandwidth per user	Relatively slow rollout
		Poor customer experience during installation and service
LMDS/MMDS	Can offer broadband to areas otherwise not provided "wired" access	Subscribers must be within line of sight
	Offers higher data rates	High weather and radio interference
		Difficult to gain roof rights
		Lack of standards for Interoperability
		Higher Installation cost
Fiber to the home	Very high bandwidths and thus excellent video/TV quality	High investment and deployment costs
	Simple network architecture	Not trivial to move connections
		Hard to reach rural areas

Table 5. Broadband Access Technologies

- DSL is typically billed at a flat rate, so you can use it as much as you want without incurring more charges.
- There is no dialup—DSL makes the Internet available 24×7 (24 h/day, 7 days/week).
- Phone company networks are among the most reliable, with only minutes of downtime each year.
- Because the connection is made over individual phone lines, each user has a point-to-point connection to the Internet.

5.2.1.2. Types of DSL. Table 6 shows the characteristics of the DSL technologies. DSL is sometimes referred to as xDSL because there are several different variations:

- Asymmetric DSL (ADSL)—delivers high-speed data and voice service over the same line. The distance from the CO determines speeds; as the distance increases, the speed available decreases.
- *G.Lite*—variation of ADSL, a DSL that the end user can install and configure. It is not yet fully plug-and-play, and has lower speeds than full-rate ADSL.
- Symmetric DSL (SDSL)—downstream speed is the same as upstream. Does not support voice connections on the same line. The distance from the CO

determines speeds; as the distance increases, the speed available decreases.

- *ISDN DSL* (IDSL)—a hybrid of ISDN and DSL; it's always an alternative to dialup ISDN. Does not support voice connections on the same line.
- *High-bit-rate DSL* (HDSL)—the DSL that is already widely used for T1 lines. Requires 4 wires instead of the standard single pair.
- *Very high-bit-rate DSL* (VDSL)—still in an experimental phase, this is the fastest DSL, but deliverable over a short distance from the CO.
- *Voiceover DSL* (VoDSL) an emerging technology that allows multiple phone lines to be transmitted over one phone wire, while still supporting data transmission. VoDSL can be used for small businesses that can balance a need for several phone extensions against their Internet connectivity needs.

Figure 20 shows an example of the ADSL network architecture.

5.2.2. Cable Access. Cable systems were originally designed to deliver broadcast television signals efficiently

Table 6. DSL Technologies

DSL Service	Data Speeds	Information
ADSL (asynchronous DSL)	Downstream: 1.5–1.8 Mbps	Most popular DSL service — based on inherent traffic flow of Internet
	Upstream: 64 kbps	Operating range <18,000 ft from CO
		Speeds are faster when close to CO
		Well suited for high-speed Internet/intranet access and telecommuter applications
ADSL Lite	Downstream: 1 Mbps	Operating range up to 18,000 ft from CO
	Upstream: 384 Kbps	Can travel over longer distances than most DSL services
		Does not require a splitter, thus limiting installation and service costs
		Consumer Internet access
SDSL (synchronous DSL)	160 kbps-2.3 Mbps	Uses one line
		Operating range <10,000 ft from CO
		Suited for videoconferencing applications and/or remote LAN access
		Business Internet access
IDSL (ISDN DSL)	144 kbps downstream and upstream	Operating range up to 18,000 ft from CO (extra equipment can increase distance)
HDSL (high-bit-rate DSL)	1.5 Mbps downstream and upstream	Uses two or four lines
		Operating range <12,000 ft from CO
		Replaces T1/E1 service
		Used primarily for PBX network connections, Internet servers, and private data networks
VDSL (very-high-bit-rate DSL)	26 Mbps symmetric	Symmetric and asymmetric configurations
	${<}52$ Mbps asymmetric downstream	High-capacity service usually served to SOHO/SME users
	< 6.4 Mbps asymmetric upstream	Capable of HDTV delivery
		Operating range 1000–4500 ft from CO
		Positioned as service of choice for eventual fiber-based all-optical networks

to subscribers' homes. The coaxial cable systems typically operate with 330 or 450 MHz of capacity, whereas hybrid fiber-coax (HFC) systems are expanded to \geq 750 MHz.

Logically, downstream video programming signals begin around 50 MHz, the equivalent of channel 2 for over-the-air television signals. The 5-42 MHz portion of the spectrum is usually reserved for upstream communications from subscribers' homes. Each standard television channel occupies 6 MHz of RF the spectrum. Thus a traditional cable system with 400 MHz of downstream bandwidth can carry the equivalent of 60 analog TV channels and a modern HFC system with 700 MHz of downstream bandwidth has the capacity for some 110 channels.

To support data services over a cable network, a television channel, in the 50-750-MHz range, is typically allocated for downstream traffic to homes and other channels, in the 5-42-MHz band are used to carry upstream signals.

A head-end cable modem termination system (CMTS) communicates through these channels with cable modems located in subscriber homes to create a virtual local-area network (LAN) connection. Most cable modems are external devices that connect to a personal computer through a standard 10base-T Ethernet card or universal serial bus (USB) connection, although internal PCI modem cards are also available. The cable modem access network operates at layer 1 (physical) and layer 2 (media access control/logical link control) of the OSI Reference Model. Thus, layer 3 (network) protocols, such as IP traffic, can be seamlessly delivered over the cable modem, platform to end users.

A single downstream 6-MHz television channel may support up to 27 MHz of downstream data throughput from the cable head end using 64-QAM (quadrature amplitude modulation) transmission technology. Speeds can be boosted to 36 Mbps using 256-QAM. Upstream channels may deliver 500 kbps-10 Mbps from homes



Figure 20. ADSL network architecture (using a splitter).

using 16-QAM or QPSK (quadrature phase shift key) modulation techniques, depending on the amount of spectrum allocated for service. This upstream and downstream bandwidth is shared by the active data subscribers connected to a given cable network segment, typically 500–2000 homes on a modern HFC network.

An individual cable modem subscriber may experience access speeds from 500 kbps to 1.5 Mbps or more—depending on the network architecture and traffic load—blazing performance compared to dialup alternatives. However, when surfing the Web, performance can be affected by Internet backbone congestion.

5.2.2.1. Data over Cable Service Interface Specification (DOCSIS). The DOCSIS was developed by the North American Cable Industry under the auspices of Cable Labs to create a competitive market for cable modem equipment. It was developed as a cheap Web-serving platform. The

main specification work for DOCSIS 1.0 was completed in March 1997 [27].

A cable data system consists of multiple cable modems (CMs), in subscriber locations, and a cable modem termination system (CMTS), all connected by a CATV plant. The CMTS can reside in a head-end or a distribution hub. DOCSIS products have been available since 1999. The DOCSIS 1.1 version has enhanced the specification in terms of quality of service (QoS), IP multicast, and security. The DOCSIS 2.0 version has been released in 2002 downstream. The DOCSIS supports an upstream of 320 kbps-10.24 Mbps and downstream rates of 36 Mbps.

5.2.2.2. DOCSIS Cable Modem Protocol. Figure 21 shows the cable modem protocol stack. CM performs the lower four layers. It receives IP over Ethernet, adds encryption, mediates access to the return path, and modulates the data on to the cable network on the forward



Figure 21. DOCSIS cable modem protocol stack.

path. CMTS also adds an MPEG-2 framing layer. Above the DOCSIS protocol layers is the IP layer and then the Internet applications and protocols.

5.2.3. Fixed Wireless Access. New wireless technologies such as LMDS, MMDS, and 38-GHz platforms provide very high data rates to end users. Many wireless networks utilize hybrids of wireless, satellite, fiber, and copper to create a complete end-to-end platform [20,28].

The traditional version of fixed wireless solution for point-to-point uses microwave at 1.7–40 GHz. The 38-GHz carrier service includes 14 pairs of 50-MHz wide channels available for last-mile connections. Spectrum is licensed primarily to Winstar and Advanced Radio Telecommunications. The technology is mainly used to extend fiber networks.

5.2.3.1. Local Multipoint Distribution Service (LMDS). The broadband LMDS is used to provide point-to-multipoint communications, two-way voice data, Internet access, and video services. It operates within 27.3–31.5 GHz. Cells are no larger than 2 mi in radius but require many cells to cover a large area. LMDS can serve roughly 80,000 customers from a single node. The downstream data rates top out at 38 Mbps.

5.2.3.2. Multipoint Distribution System (MMDS). MMDS is based on 200 MHz of spectrum allocated for TV transmission, increasing the flexibility for two-way communications. It allows transmission up to 1 Gbps per band for send and receive applications within a 35-mi radius. MMDS utilizes cellularization and is well suited for residential and small-business markets.

5.3. Satellite Access Networks

The access systems provide regional coverage and are less complicated than their global connected counterparts. They are more cost-effective, have less associated technical risk, and have less regulatory issues [29]. Table 7 provides a partial list of access satellite systems for regional coverage.

5.3.1. DVB Satellite Access. In this section Internet access via satellite using digital videobroadcasting-return channel by satellite (DVB-RCS) is discussed. The DVB network elements consist of an enterprise model, servicelevel agreements (SLA), and TCP Protocol Enhancement Proxy (PEP), the hub station, and the satellite interactive terminal (SIT). The target applications of the DVB network could be small and medium enterprises and residential users. One of the major advantages of DVB-RCS is that multicasting is possible at a low cost using the existing Internet standards. The multicast data is tunneled over the Internet via a multicast streaming feeder link from a streaming source to a centralized multicast streaming server and is then broadcasted over the satellite medium to the intended target destination group. The DVB-RCS system supports relatively large streaming bandwidths compared to existing terrestrial solutions (from 64 Kbps to 1 Mbps).

In the DVB network, a satellite forward and return links typically use frequency bands in Ku (12-18 GHz) and/or Ka (18-30 GHz). The return links use spot beams and the forward link global beams are used for broadcasting and Ku band. Depending on the frequency bands (Tx/Rx), three popular versions are available: (1) Ku/Ku (14/12 GHz), (2) Ka/Ku (30/12 GHz), and (3) Ka/Ka (30/20 GHz). In business-to-business applications, the SIT is connected to several user PCs via a LAN and a point-of-presence (POP) router. The hub station implements the forward link via a conventional DVB-S chain (similar to Digital TV broadcasting) whereby the IP packet is encapsulated into DVB streams, IP over DVB. The return link is implemented using the DVB-RCS standard MF-TDMA Burst Demodulator Bank, IP over ATM. The hub station is connected to the routers of several ISPs via a broadband access server. The hub maps the traffic of all SITs belonging to each ISP in an efficient way over the satellite. The selection of a suitable residential access technology depends on the type of application, site location, required speed, and affordable cost.

5.3.1.1. DVB-RCS. The DVB return channel system via satellite (DVB-RCS) was specified by an ad hoc ETSI

Services	StarBand	WildBlue	iPStar	Astra-BBI	Cyberstar
Data uplink	38–153 kbps	384 kbps-6 Mbps	2 Mbps	2 Mbps	0.5-6 Mbps
Data downlink	40 Mbps	384 kbps-20 Mbps	10 Mbps	38 Mbps	Max. 27 Mbps
Coverage area	USA	Americas	Asia	Europe	Multiregional
Market	Consumer	Business/SME	Consumer and business	Business	ISPs, multicast
Terminal cost (U.S.\$)	<\$350	<\$1000	<\$1000	\sim \$1800 <\$450 (2001)	—
Monthly Access fee (U.S.\$) Antenna size (M)	\$60 1.2	$$45 \\ 0.8-1.2$	0.8–1.2		_
Frequency band	Ku	Ka	Ku, Ka	Ku/Ka	Ku, Ka
Satellite	GEO	GEO	GEO	GEO	GEO
Operation scheduled	Nov. 2000	Mid-2002	Late 2002	Late 2000	1999 - 2001

 Table 7. Broadband Access Systems

technical group founded in 1999. The DVB-RCS system specification in ETSI EN 301 790, v1.2.2 (2000–2012) specifies a satellite terminal (sometimes known as a *satellite interactive terminal* (SIT) or *return channel satellite terminal* (RCST) supporting a two-way DVB satellite system [30,31]. Another CDMA-based spread ALOHA has been proposed for return channel access [32]. This section describes the DVB-RCS protocol. The use of standard system components provides a simple approach and should reduce time to market.

Customer premises equipment (CPE) receives a standard DVB-S transmission generated by a satellite gateway. Packet data may be sent over this forward link in the usual way (e.g., MPE, data streaming) DVB-RCS provides transmit capability from the user site via the same antenna. The transmit capability uses a multifrequency time-division multiple access (MFTDMA) access scheme to share the capacity available for transmission by the user terminal. The return channel is coded using rate $\frac{1}{2}$ convolution FEC and Reed-Solomon coding. The standard is designed to be frequency-independent and does not specify the frequency band(s) to be used — thereby allowing a wide variety of systems to be constructed. Data to be transported may be encapsulated in ATM cells, using ATM adaptation layer 5 (AAL-5), or use a native IP encapsulation over MPEG-2 transport. It also includes a number of security mechanisms.

Figure 22 shows an example of broadband satellite network using the DVB-RCS standard for the return channel protocol.

5.3.1.2. DVB-RCS-CPE Operations. A return channel satellite terminal (RCST), once powered on, will start to receive general network information from the DVB-RCS network control center (NCC). The NCC provides monitoring and control functions, and generates the control and timing messages required for operation of the satellite network. All messages from the NCC are sent using the MPEG-2 TS using private data sections

(DVB SI tables). These are transmitted over the forward link. Actually the DVB-RCS specification calls for two forward links—one for interaction control and another for data transmission. Both links can be provided using the same DVB-S transport multiplex. The term "forward link" refers to the link from the gateway that is received by the user terminal. DVB-RCS allows this communication to use the same transmission path as sued for data (i.e., the DVB-S receive path), or an alternative interaction path. Conversely, the return link is the link from the user terminal to the gateway using the DVB interaction channel. The control messages received over the forward link also provide the network clock reference (NCR).

The NCC controls user terminal transmissions. Before a terminal can send data, it must first join the network by communicating (logging on) with the NCC describing the configuration. The logon message is sent using a frequency channel also specified in the control messages. This channel is shared between user terminals wishing to join the network using the slotted ALOHA access protocol. After receiving a logon message from a valid terminal, the NCC returns a series of tables including the terminal burst time plan (TBTP) for the use terminal. The MF-TDMA burst time plan (TBTP) allows the terminal to communicate at specific time intervals using specific assigned carrier frequencies at an assigned transmit power.

The terminal transmits a group of ATM cells (or MPEG-TS packets). This block of information may be encoded in one of several ways using convolutional coding, RS/convolutional coding or Turbo coding. The block is prefixed by a preamble and optional control data and followed by a postamble to flush the convolutional encoder. The complete burst is sent using QPSK modulation. Before each terminal can use its allocated capacity, it must first achieve physical-layer synchronization of time, power, and frequency, a process completed with the assistance of special synchronization messages sent over the satellite channel. A terminal normally logs off the system when it



Figure 22. Broadband satellite access - DBV-RCS.



Figure 23. Future mobile wireless technologies.

has completed its communication. Alternately, if there is a need, the NCC may force a terminal to log off.

5.4. Mobile Wireless Access

Figure 23 shows a future trend of mobile communication systems. The major requirements to be supported by these systems are high data rate, high mobility, and seamless coverage. It might be difficult to realize a single system satisfying all these requirements. Some of them can provide high data rates, and others can support high mobility and coverage. However, the future intelligent integrated system solutions could satisfy all the user demands. The first-generation systems are analog cordless and analog cellular. The second-generation systems are digital systems with digital cellular such as GSM, IS54 digital cellular, personal digital cellular (PDC), and IS95. The digital cordless are DECT, PHS. These systems are operated nation wide or internationally and are today's mainstream ones. The data rates for users in air links are limited to less than several tens of kilobits per second. The IMT-2000 is the third-generation (3G) cellular systems, which provide 2 Mbps and 144 kbps indoor vehicular environments. Satellite-based Iridium and GLOBALSTAR belong to this category. As candidates of future mobile systems, 4G cellular support high data rates up to 20 Mbps and high mobility include intelligent transport systems and high-altitude stratospheric platform stations (HAPSs) [33-37].

6. STANDARDS STATUS

The standardization process is in progress for the various broadband access technologies discussed in the previous

Table 8. Technology Standards and Organizations

Technology	Standard/Organization			
IP	IETF (<i>http://www.ietf.org</i>)			
	ITU-T (<i>http://www.itu.int/ITU-T</i>)			
	MPLS Forum (http://www.mplsforum.org)			
FR	Frame Relay Forum (http://www.frforum.com)			
ATM	ATM Forum (http://www.atmforum.com)			
Cable	Cable Labs (http://www.cablelabs.com)			
DSL	DSL Forum (<i>http://www.adsl.com</i>)			
Wireless	DAVIC (http://www.davic.org)			
Video	MPEG (http://www.mpeg.org)			
	DVB (<i>http://www.dvb.org</i>)			
	ETSI (http://www.etsi.org)			
Broadband content delivery	BCD (http://www.bcdforum.org)			
Satellite ATM	ITU-R (http://www.itu.int/ITU-R)			
Satellite IP	ITU-R (<i>http://www.itu.int/ITU-R</i>)			
	ITU-T (<i>http://www.itu.int/ITU-T</i>)			
	IETF (<i>http://www.ietf.org</i>)			
	TIA (http://www.tiaonline.org)			
DVB-RCS	ETSI/DVB-RCS (http://www.etsi.org)			

section. The different standards organizations and the fora along with their Websites are provided in Table 8.

7. FUTURE NETWORKING: CHALLENGES

The broadband network infrastructure for the enterprise and residential access, for the emerging applications such as videostreaming, content distribution delivery, telemedicine, two-way interactive learning, and games, must address the following challenges:

- High-speed access
- QoS
- Scalability
- Interworking
- Interoperability
- Security
- Cost-effective solutions (cost per bit)

Many of the standards organizations addressed in Section 6 are developing technical solutions and recommendations for interoperable infrastructure. One of the examples based on MPLS is discussed in the following paragraphs.

7.1. Multiprotocol Label Switching (MPLS)

Multiprotocol label switching (MPLS) is a switching method in which a label field in the incoming packets is used to determine the next hop [38,39]. At each hop, the incoming label is replaced by another label that is used at the next hop. The path thus realized is called a labelswitched path (LSP). Devices that base their forwarding decision solely on the basis of the incoming labels (and ports) are called *label-switched routers* (LSRs). In MPLS, the assignment of a particular packet to a particular of forwarding equivalence classes (FECs) is done just once, as the packet enters the network. The FEC to which the packet is assigned is encoded as a short fixed-length value known as a "label." When a packet is forwarded to its next hop, the label is sent along with it; that is, the packets are "labeled" before they are forwarded.

At subsequent hops, there is no further analysis of the packet's network-layer header. Rather, the label is used as an index into a table, which specifies the next hop, and a new label. The old label is replaced with the new label, and the packet is forwarded to its next hop. In the MPLS forwarding paradigm, once a packet is assigned to FEC, subsequent routers do no further header analysis; the labels drive all forwarding. This has a number of advantages over conventional network-layer forwarding and is a great tool for traffic engineering.

Traffic engineering (TE) is concerned with performance optimization of operational networks. In general, it encompasses the application of technology and scientific principles to the measurement, modeling, characterization, and control of Internet traffic, and the application of such knowledge and techniques to achieve specific performance objectives [40,41]. A major goal of Internet traffic engineering is to facilitate efficient and reliable network operations while simultaneously optimizing network resource utilization and traffic performance. Traffic engineering has become an indispensable function in many large autonomous systems because of the high cost of network assets and the commercial and competitive nature of the Internet. All these factors emphasize the need for maximal operational efficiency, which TE can help to achieve. MPLS-based traffic engineering is a good candidate to provide hard QoS guarantees to important services such as voice over IP, video and multimedia over IP, and virtual private networks [42]. MPLS can help to maximize both resource utilization and QoS offered to a given traffic or aggregation of traffics. There are efforts under way to develop satellite over MPLS in addition to MPLS over terrestrial networks to provide QoS guaranteed solutions for both wired and wireless satellite networks.

7.2. Future MPLS Enterprise Network Example

Figure 24 shows a future MPLS-based enterprise network example.

MPLS-based VPNs are widely considered to be a viable next-generation VPN technology MPLS-VPNs were developed to simplify the VPNs without requiring encryption. Instead, MPLS used tunnels to create connectivity between sites. This allows enterprises to nail up bandwidth between sites, providing for bandwidth and connection reliability along with ensuring predictable service for different applications. However, one of the drawbacks is that the processing power required to support each MPLS-VPN becomes a bottleneck on existing monolithic routers because of their centralized router processor architecture that limit processing scalability.

In addition, all the core routers in the network must be configured to run Border Gateway Protocol (BGP). BGP is a major task for MPLS. To receive the full benefits of MPLS-VPNs, vendors must cooperate heavily while implementing the standards without prioritized methods. Some of the benefits of MPLS based VPNs are

- Service guarantees provided with respect to packet prioritization and bandwidth reservation
- Scalable infrastructure
- Seamless ATM interworking
- Small to large group connectivity

But some of the shortcomings of MPLS-VPNs are

- The inability of the enterprise to manage its network independent of the carrier
- Inability of scaling of the MPLS edge control plane



Figure 24. MPLS-based enterprise network.

• Scaling the number of virtual routers within a monolithic system

However, MPLS with traffic engineering developments provide quality of service (QoS) of the network.

8. CONCLUSIONS

Increasing demand for high-speed applications over Internet is driving new broadband network infrastructures. The applications range from simple file transfer and remote login to IP multicast, media streaming, and content delivery distribution. These emerging applications require larger bandwidths than 64 kbps or T-1 rates, and user service guarantees as opposed to "best effort" service over the today's public Internet. To meet such application requirements, different technology options are available. However, research and development in the areas of efficient protocols, QoS architectures, and security mechanisms is urgently required.

In this article, enterprise access technologies such as Gigabit Ethernet, frame relay, ATM, IP and broadband satellite technologies with network examples were provided. The broadband access technologies for residential access, DSL, cable, hybrid coax-fiber, and satellite were also discussed. Current system examples were provided. These discussions are in no way completely exhaustive. The references should provide additional resources for more depth on any single technology topic.

Finally, future networking requirements including speed, QoS, interoperability, security, and cost per bit were described with an illustrative example of an MPLS based enterprise solution.

ACRONYMS

10GMII	10-gigabit media-independent interface				
AAL	ATM adaptation layer				
ABR	Available bit rate				
ACTS	Advanced communication technology				
	satellite				
ADSL	Asymmetric DSL				
AH	Application header				
ARPANET	Advanced Research Projects Agency				
	Network				
ATM	Asynchronous transfer mode				
BCDF	Broadband Content Delivery Forum				
BECN	Backward error congestion notification				
BGP	Border Gateway Protocol				
B-ISDN	Broadband Integrated Service Data				
	Network				
B2B	Business to business				
CAC	Connection admission control				
CBR	Constant bit rate				
CDD	Content delivery distribution				
CDV	Cell delay variation				
CDVT	Cell delay variation tolerance				
CLP	Cell loss priority				
CLR	Cell loss ratio				
CM	Cable modem				
CMTS	Cable modem termination system				

CPE	Customer premise equipment
\mathbf{CR}	Command response
CTD	Cell transfer delay
DAMA	Demand assignment multiple access
DAVIC	Digital Audio-Visual Council
DE	Discard eligibility
DLCI	Data-link control identifier
DOCSIS	Data over Cable Service Interface
	Specification
DS-1	Digital signal level one
DSL	Digital subscriber line
DSLAM	Digital subscriber line access multiplexer
DVB	Digital video broadcast
DVB-S	Digital video broadcast-satellite
DVB-RCS	Digital video broadcast–return channel system
DWDM	Dense wave-division multiplexing
EA	Extended address
EH	Ethernet header
ENIAC	Electronic numerical integrator and
	computer
ETSI	European Telecommunication Standards
1101	Institute
FCS	Frame check sequence
FEC	Forward error correction: Forwarding
120	equivalence class
FECN	Forward explicit congestion notification
FR	Frame relay
FSS	Fixed satellite service
FTP	File Transfer Protocol
GEO	Geostationary Earth Orbit
GigE	Gigabit Ethernet
GFC	Generic flow control
GFR	Guaranteed frame rate
GPRS	General packet radio service
GSM	Global System for Mobile Communication
GSO	Geosynchronous orbit
GW	Gateway
HAPS	High-altitude stratospheric platform
	station
HDTV	High-definition TV
HDSL	High-bit-rate DSL
HEC	Header error control
HFC	Hybrid fiber-coaxial cable (coax)
HTTP	Hyper Text Transfer Protocol
ICMP	Internet Control Message Protocol
IDSL	ISDN DSL
IEEE	Institute of Electronics and Electrical Engineers
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
IMT2000	International Mobile Telecommunications
IP	Internet Protocol
IPv4/IPv6	Internet Protocol version 4/Internet
	Protocol version 6
IPH	IP header
IPSec	Internet Protocol Security
ISDN	Integrated Services Digital Network
ISP	Internet service provider
ITU-R	International Telecommunications
	Union — Kadio Sector

2676 TRENDS IN BROADBAND COMMUNICATION NETWORKS

ITU-T	International Telecommunications
	Union—Telecommunications
LAN	Local-area Networks
Lec	Local Exchange Carrier
LMDS	Local multipoint distribution services
LSP	Label-switched plan
LSR	Label-switched router
MAC	Media access control
MAN	Metropolitan-area network
MBS	Maximum burst size
MCR	Minimum cell rate
MF-TDMA	Multifrequency time-division multiple access
MMDS	Multichannel multipoint distribution services
MPE	Multiprotocol encapsulation
MPEG	Moving Picture Expert Group
MPLS	Multiprotocol label switching
MSS	Mobile satellite service
MTU	Maximum transmission unit
NCC	Network control center
NCR	Network clock reference
NCS	Network control station
NGSO	Nongoogynghronous orbit
NNTD	Notwork News Transfer Protocol
nn IF	New week time werichle bit rete
	Optical commism
OC	Optical carrier
USI	Deinste hausele auch
PBA	Private branch exchange
PCR	Peak cell rate
PDC	Personal digital cellular
PEP	Protocol enhancement proxy
PMA	Physical medium attachment
PMD	Physical medium-dependent
POP	Point of presence
PSTN	Public switched telephone network
PT	Payload type
PVC	Permanent virtual circuit
QAM	Quadrature amplitude modulation
QoS	Quality of service
QPSK	Quadrature phase shift key
RCST	Return channel satellite terminal
RFC	Request for Comments (IETF Document)
rt-VBR	Real-time variable bit rate
SCR	Sustainable cell rate
SDSL	Symmetric DSL
SIT	Satellite interactive terminal
SLA	Service-level agreement
SME	Small- and medium-size enterprises
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SOHO	Small office/home office
SONET	Synchronous ontical network
SS7	Signaling System 7
SVC	Switched virtual circuits
ТВТР	Terminal hurst time plan
	Transmission Control Protocol
	Transmission Control Protocol Harden
	Time division welt-planing
	I ime-division multiplexing
UBK	Unspecified bit rate
UDP	User Datagram Protocol

UMTS	Universal Mobile Telecommunication
	System
UNI	User-network interface
UPC	Usage parameter control
VCI	Virtual channel identifier
VDSL	Very-high-bit-rate DSL
VOD	Video on demand
VoDSL	Voice over DSL
VPI	Virtual path identifier
VPN	Virtual private network
WAN	Wide-area network
WDM	Wavelength-division multiplexing

BIOGRAPHY

Sastri Kota has been a technical consultant with Loral Skynet, Palo Alto, California since 2001. Since the early 1970s he has held various technical and management positions and contributed to the military and commercial satellite systems in the areas of network design, broadband ATM and IP network architectures, and protocol analyses at Lockheed Martin, SRI International, Ford Aerospace, The MITRE and Computer Sciences Corp. Currently he is the U.S. Chair for ITU-R, Working Party 4B. He was the Chair for Wireless ATM Working Group and was the recipient of the ATM Forum Spotlight award. He holds a B.S in Physics, B.S.E.E, M.S.E.E from India, and Electrical Engineer's degree from Northeastern University, Boston. He has published over 90 technical papers in journals, book chapters, and conference proceedings. He was the Guest Editor for IEEE Communications Magazine. He has served as Satellite Communications Symposium Chair for IEEE GLOBECOM '00, Assistant Technical Chair for IEEE MILCOM '97, '90, and SPIE '91 conferences. He also served on conference technical committees and as Session Chair for IEEE GLOBECOM, ICC, WCNC, MILCOM, and AIAA communication satellite systems, in the areas of broadband, wireless, and satellite networks. His research interests include QoS for satellite IP, traffic management, wireless and mobile IP networks, and broadband access. He is a senior member of IEEE, Associate Fellow of AIAA, and member of ACM.

BIBLIOGRAPHY

- L. G. Roberts and B. D. Wessler, Computer network development to achieve resource sharing, *Proc. AFIPS Conf.* 1970, Vol. 36, pp. 543–549.
- L. Kleinrock, Queuing Systems, Vol. 2; Computer Applications, Wiley, New York, 1976.
- 3. Renters Story Network Fusion, April 2002.
- S. Aidarous, Challenges in evolving to IP-based global networks, *GLOBECOM 2001*, San Antonio, TX, Nov. 25–29, 2002.
- S. Kota, Quality of Service (QoS) Architecture for Satellite IP Networks, Document 4B/86, ITU-R Working Party 4B, Geneva, Switzerland, 2002.
- J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2000.

TRENDS IN WIRELESS INDOOR NETWORKS 2677

- R. Ramaswami and K. Sivarajan, Optical Networks: A Practical Perspective, Academic Press/Morgan Kaufman, 1998.
- 8. S. Saunders, *Data Communications Gigabit Ethernet Handbook*, McGraw-Hill, New York, 1998.
- P. Smith, Frame Relay, Principles and Applications, Addison-Wesley, Reading, MA, 1993.
- 10. Frame Relay Forum, http://www.frforum.com.
- H. J. R. Dutton and P. Lenhard, Asynchronous Transfer Mode (ATM): Technical Overview, Prentice-Hall, Saddle River, NJ, 1995.
- D. E. McDysan and D. L. Spohn, ATM Theory and Application, McGraw-Hill Series on Computer Communications, McGraw-Hill, New York, 1998.
- 13. The ATM Forum, *Traffic Management Specification*, version 4.0, 1996.
- W. R. Stevens, TCP/IP Illustrated, Vol. 1, The Protocols, Addison-Wesley, Reading, MA, 1994.
- S. Deering and R. Hinden, *Internet Protocol Version 6* (IPv6) Specification, IETF, RFC 2460, Dec. 1998.
- S. Kent and R. Atkinson, Security Architecture for the Internet Protocol, Internet RFC 2401, 1998.
- G. Maral and M. Bousquet, Satellite Communications Systems, Systems, Techniques and Technology, Wiley, New York, 1993.
- S. Kota, A. Durresi, and R. Jain, Realizing future broadband satellite network services, in *Modeling and Simulation Environment for Satellite and Terrestrial Communication Networks*, A. Nejat Ince, ed., Kluwer, Boston, 2002.
- Pioneer Consulting, Broadband Satellite: Analysis of Global Market Opportunities and Innovation Challenges, 2002.
- R. J. Bates, Broadband Telecommunications Handbook, Mc-Graw-Hill, New York, 2000.
- R. W. Smith, Broadband Internet Connections: A User's Guide to DSL and Cable, Addison-Wesley, Reading, MA, 2002.
- 22. W. J. Goralski, ADSL, McGraw-Hill, New York, 1998.
- 23. G. Abe, Residential Broadband, CISCO Press, 1997.
- I. Cooper and M. A. Bramhall, ATM passive optical networks and integrated VDSL, *IEEE Commun. Mag.* 38(3): 174–179 (2000).
- 25. A. Azzam, *High-Speed Cable Modems*, McGraw-Hill, New York, 1997.
- 26. G. Held, Next-Generation Modems: A Professional Guide to DSL and Cable Modems, Wiley, New York, 2000.
- D. Fellows and D. Jones, DOCSIS™ cable modem technology, IEEE Commun. Mag. 39(3): 202–209 (2001).
- 28. J. R. Vacca, Wireless Broadband Networks Handbook, McGraw-Hill, New York, 2001.
- A. Durresi and S. Kota, Satellite TCP/IP, in *High Performance TCP/IP*, M. Hassan and R. Jain, eds., Prentice-Hall, Englewood Cliffs, NJ, 2002.
- DVB, Interactive Channel for Satellite Distribution Systems, DVBRCS001, rev. 14, ETSI EN 301 790, V1.22 (2000–2012).
- J. Neale, R. Green, and J. Landovskis, Interactive channel for multimedia satellite networks, *IEEE Commun. Mag.* 39(3): 192–198 (2001).

- 32. S. Kota, M. Vazquez-Castro, and J.Carlin, Spread ALOHA multiple access for broadband satellite return channel, Proc. AIAA 20th Int. Communication Satellite Systems Conf., 2002.
- 33. S. Ohmori, Y. Yamao, and N. Nakayima, The future generations of mobile communications based on broadband access methods, *Int. J. Wireless Pers. Commun.* 17(2–3): 175–190 (2001).
- 34. C. Perkins, IP Mobility Support, Internet RFC 2002, 1996.
- J. Rapeli, Future directions for mobile communications business, technology and research, Int. J. Wireless Pers. Commun. 17(2-3): 155-173 (2001).
- 36. K. Pahlavan and P. Krishnamurthy, Principles of Wireless Networks—a Unified Approach, Prentice-Hall, Englewood Cliffs, NJ, 2002.
- G. Wu, M. Mizuno, and P. J. M. Havinga, MIRAI architecture for heterogeneous network, *IEEE Commun. Mag.* 40(2): 126-134 (2002).
- A. Ghanwani et al., Traffic engineering standards in IP network using MPLS, *IEEE Commun. Mag.* 37(12): 49-53 (1999).
- G. Swallow, MPLS advantages for traffic engineering, *IEEE Commun. Mag.* 37(12): 54–57 (1999).
- E. W. Gray, MPLS: Implementing Technology, Addison-Wesley, New York, 2001.
- B. Davie and Y. Rekhter, MPLS Technology and Applications, Morgan Kaufman, San Francisco, 2000.
- 42. G. Armitage, Quality of Service in IP Networks, Foundations for a Multi-Service Internet, MTP, Indianapolis, IN, 2000.

TRENDS IN WIRELESS INDOOR NETWORKS

K. PAHLAVAN J. BENEAT X. LI Center for Wireless Information Network Studies Worcester Polytechnic Institute Worcester, Massachusetts

1. INTRODUCTION

This article provides an overview of the trends in wireless indoor networks. Since the late 1970s, when the concept of wireless LAN was first introduced, this area has gone through significant ups and downs. After a very exciting development period in the late 1980s, market revenues still remained far below predictions in the mid-1990s. However, during the late 1990s with the introduction of wireless personal area networking, Bluetooth technology, and home networking, a new surge of excitement has emerged in the wireless indoor communication industry, resulting in a number of new startup projects. Furthermore, the emergence of wireless indoor positioning systems to augment wireless indoor telecommunication services has attracted tremendous attention to wireless indoor networks. This chapter starts with an overview of the traditional wireless LAN industry and then moves to explain the new emerging wireless personal-area network (WPAN), home networking, and indoor geolocation industries.

2. WIRELESS LANs

Since 1980 the perception of a WLAN industry has evolved. It was implemented on a variety of innovative technologies and raised great hopes for developing a sizable market a couple of times. Today, the major differentiation of WLANs from wide area cellular services is the method of delivery to the users, data rate limitations, and frequency band regulations. Cellular data services are delivered by operating companies as services while WLAN users own their network. At a time where the 3G cellular industry is striving for 2-Mbps (megabit/second) packet data services, WLAN standards are focusing on 54-Mbps services. Another differentiation with other radio networks is that, today, almost all WLANs operate in unlicensed bands where frequency regulations are loose and there is no charge or waiting time to obtain the band. To obtain a deeper understanding of all these issues, it is very useful to go over the history of the WLAN industry to see how all these unique issues evolved.

2.1. Early Experiences

The idea of a wireless LAN was first introduced by Gfeller at the IBM Rueschlikon Laboratories in Switzerland in the late 1970s [1]. The number of terminals in manufacturing floors was growing and wiring within the manufacturing floor was difficult. In offices, wires are normally snaked under the suspended ceilings and through the interior partitioning wall. This was not possible in manufacturing floors. In offices, in extreme cases, the wiring could be installed under the floor or simply laid on the floor with some cover. In manufacturing floors, the floors are more rugged, making underfloor wiring more expensive, while throwing wires over the floor is not acceptable because of the danger presented by moving heavy machinery. Diffused IR technology was selected for the implementation of the WLAN. Diffused IR avoided interference problems due to electromagnetic signals radiating from the machinery and avoided dealing with cumbersome administrative procedures from frequency administration agencies. Unfortunately, the principal researcher abandoned the project when the goal of 1 Mbps with reasonable coverage did not materialize.

At about the same time, a second noticeable project on WLANs was performed by Ferert at Hewlett-Packard Pal Alto Research Laboratories in California [2]. In this project, a 100-kbps direct-sequence spreadspectrum (DSSS) WLAN operating at 900 MHz using a carrier sense multiple-access (CSMA) access method technique was developed for office areas. The project was conducted under an experimental license agreement from the FCC. However, when Ferert filed to obtain bands from the FCC, he was discouraged by the administrative complexity of securing a band for his application and he also abandoned the project. A couple of years later, Codex/Motorola attempted to implement a WLAN at 1.73 GHz, but the project was also dropped during negotiations with the FCC. Although all the pioneering WLAN projects were abandoned, the area continued to attract attention and negotiations with the FCC to secure bands continued [3]. These projects revealed several important challenges facing the WLAN industry that still remain:

- 1. Complexity and cost WLAN implementation alternatives using IR, spread-spectrum communication, or traditional radios are far more complex and diversified than those in wired LANs.
- 2. *Bandwidth*—data rate limitations in a wireless medium are far greater than in a wired medium.
- 3. Coverage point-to-point coverage of a WLAN operating in a building is smaller than cables or even twisted-pair (TP) LAN solutions.
- 4. *Interference* WLANs are subject to interference from other overlaying WLANs or other devices operating in the same frequency medium.
- 5. *Frequency* administration radiofrequency-based WLANs are subject to expensive and timely frequency regulations.

2.2. Emergence of Unlicensed Bands

Wireless LANs need a significant amount of bandwidth, at least several tens of MHz. Yet, in the 1980s it had not been shown that a strong market existed, comparable to that of the cellular voice industry when it originally started with two 25-MHz bands. In addition, frequency bands of comparable bandwidth for PCS applications were auctioned in the United States for tens of billions of dollars while the WLAN market was under a billion dollars per year. The dilemma for the frequency administration agencies was how to justify this frequency allocation. In the mid-1980s, the FCC found two solutions to this problem. The first and simplest solution was to go beyond the 1-2-GHz band used for cellular telephone and PCS applications to higher frequencies at several tens of GHz where plenty of unused bands were available. This solution was first negotiated between Motorola and the FCC, resulting in Altair, the first wireless LAN product operating in a licensed 18-19-GHz band. Motorola also established a headquarter to facilitate negotiations with the FCC regarding the usage of WLANs in different locations. If the location of operation of a WLAN was substantially changed (e.g., from one town to another), those responsible for the network would contract Motorola to manage the necessary frequency administration issues with the FCC.

The second solution used a more innovative approach to the problem by resorting to the creation of unlicensed bands. In response to the pressing need for suitable bands and motivated by recent studies depicting various implementations of wireless LANs [3], Mike Marcus of the FCC initiated the release of the unlicensed ISM bands (902–928/2400–2483/5725–5875 MHz) in May 1985 [4]. The ISM bands were the first unlicensed bands for consumer product development and played a major role in the development of the WLAN industry. In simple words, licensed and unlicensed bands are compared to backyard and public gardens. Anyone who can afford it can own a private backyard (licensed band) and arrange a barbeque dinner (a wireless product). If one cannot afford to buy a house with a backyard, he/she simply moves the barbeque party to the public park (unlicensed band) where he/she should observe certain rules or etiquette that allows others to share the public resource as well. The rules enforced on ISM bands restricted the transmission power to 1 W. Modems radiating more than 1 mW had to employ spread-spectrum technology. It was perceived that spread-spectrum communication would limit interference and allow the coexistence of several wireless applications in the same band.

Encouraged by the FCC ruling [4] and some visionary publications in wireless office information networks [3,5,6], a number of WLAN product development projects mushroomed in the North American continent. By late 1980s, the first generation of WLAN products appeared in the market. These products used three different technologies: microwave technology in the licensed 18-19-GHz band, spread spectrum in the ISM bands, and IR. They were shoebox-size access points and receiver boxes. The perception at that time was that a WLAN would be used to connect workstations to the LAN wherever wiring difficulties justified using a more expensive wireless solution. Today, we call this application LAN extension [7,8]. At that time, market predictions were estimating a shift of around 15% of the LAN market to WLAN that would generate a few billion dollars of sales per year by the early the 1990s.

Also in the late 1980s, a standardization activity was initiated under the IEEE 802.4L to provide some guidance for development of WLANs. This activity soon turned to the IEEE 802.11 standard, but it would take many years, up until 1997, for the standard to be finalized. In May 1991, to create a scientific forum for the exchange of knowledge on WLANs, the first IEEE sponsored WLAN workshop was organized concurrent to the 802.11 meeting at Worcester Massachusetts [9].

In 1992, following the momentum for WLAN developments, an industrial alliance led by Apple Computers called WINForum was formed aiming at obtaining more unlicensed bands from the FCC for the so-called Data-PCS activities. WINForum finally succeeded in securing 20 MHz of bandwidth in the PCS bands that was divided into two 10-MHz bands, the 1910–1920 MHz band for *isochronous* (voicelike) and the 1920–1930-MHz band for *asynchronous* (data type) applications. The original aim of the WINForum was to secure 40 MHz

for Asynchronous applications. WINForum defined a set of rules or etiquettes for these bands to allow coexistence. There are three basic rules: (1) to listen before talk (or transmit) or LBT protocol, (2) to use low transmitter power, and (3) to restrict the duration of the transmissions. The WINForum etiquette is based on CSMA rather than CDMA spread-spectrum communications. This was a better choice since CDMA implementations require careful power control schemes that are not feasible in an uncoordinated multiuser, multivendor WLAN environment. However, spread-spectrum communications without CDMA are less bandwidth-efficient.

Another standardization activity that started in 1992 was HIPERLAN. This ETSI-based standard aimed at high-performance WLANs with data rates up to 20 Mbps, an order of magnitude higher than the original 802.11 data rates of 2 Mbps. To support these data rates, the HIPERLAN community was able to secure two 200-MHz bands at 5.15–5.35 GHz and 17.1–17.3 GHz. This initiation encouraged the FCC to release the so-called *unlicensed national information infrastructure* (U-NII) bands in 1997 at the time the original HIPERLAN now called HIPERLAN-1 was completed. Table 1 summarizes the U-NII bands and their restrictions.

Today, IEEE 802.11a and HIPERLAN-2 projects use the U-NII bands for the implementation of 54-Mbps OFDM-based WLANs. More details of the WLAN standards can be found in Section 2.5 and Table 2.

2.3. Shift in Marketing Strategy

In the first half of the 1990s, a sizable market of around a few billions of dollars per year was expected for the shoebox-type products used as LAN-extensions in indoor areas. This did not materialize. As a consequence, two new directions for product development emerged. The first and simplest approach was to take the existing shoebox-type WLAN products, boost the transmitted power to the maximum authorized limit, and add a highgain directional antenna for outdoor interbuilding LAN interconnects. This technically simple solution would allow wireless connectivity up to a few tens of kilometers with suitable rooftop antennas. The new inter-LAN wireless bridges could connect corporate LANs within range at a much lower cost than the wired alternative such as T1 carrier lines leased from PSTN service providers. The second approach was to reduce the size of the product to a PCMCIA WLAN card suitable for the laptops that

Table 1. The U-NII-bands	Table 1. The U-N	II-bands	
--------------------------	------------------	----------	--

Band of Operation (GHz)	Maximum Transmission Power (mW)	Maximum Power with Antenna Gain of 6 dBi (mW)	Maximum PSD (mW/MHz)	Applications: Suggested and/or Mandated	Other Remarks	
5.15 - 5.25	50	200	2.5	Restricted to indoor applications	Antenna must be an integral part of the device	
5.25 - 5.35	250	1000	12.5	Campus LANs	Compatible with HIPERLAN	
5.725-5.825	1000	4000	50	Community networks	Longer range in low-interference (rural) environments	

Parameters	IEEE 802.11	IEEE 802.11b	IEEE 802.11a	HIPERLAN/2	HIPERLAN/1
Status	Approved, products	Final ballot, Nov. 1999, products	In preparation	In preparation	Approved, no products
Frequency band	2.4 GHz	2.4 GHz		5 GHz	$5~\mathrm{GHz}$
PHY, modulation	DSSS: BPSK, QPSK FHSS: GFSK	DSSS: BPSK, QPSK, CCK	OFDM		GMSK
Delay spread	1, 2 Mbps: 200–400 ns		12 Mbps: 350 ns		Unknown
robustness	11 Mbps: 20–60 ns Fallback mechanism		36 Mbps: 125 ns		
Data rate	1, 2 Mbps	1, 2, 5.5, 11 Mbps	6, 9, 12, 18	8, 24, 36, 54 Mbps	23.5 Mbps
Access method	Distributed control, CSMA/CA or RTS/CTS			Central control reservation based-access, scheduled by access point	Active contention resolution, priority signalling

Table 2. Summary of WLAN Standards

were enjoying a sizable growth. However, this approach was mainly suitable for the spread spectrum products operating at lower frequencies. Figure 1 illustrates all three applications for WLANs.

The early marketing strategy used by companies for LAN-extension aimed at a horizontal market by selling individual WLAN pieces directly to the customers. In the mid-1990s, a few successful companies adopted a major shift in marketing strategy. The new strategy aimed at a vertical market by selling the entire wireless network as a complete solution. The vertical markets approached by the WLAN industry were "bar code" industries providing wireless inventory check and tracking in warehouses and manufacturing floors, financial services providing wireless financial updates in large stock exchanges, healthcare networks providing wireless mobile services inside hospitals, and wireless campus-area networks



Figure 1. Different forms of WLAN products: (a) LAN extension; (b) Inter-LAN bridge; (c) PCMCIA cards for laptops.

(WCAN) providing wireless classrooms and offices. All these efforts boosted the market for WLANs to over half a billion dollars per year during the late 1990s.

An experimental NSF-sponsored WCAN is represented in Fig. 2. It was designed as a testbed for performance monitoring of WLAN products at Center for Wireless Information Network Studies (CWINS), Worcester Polytechnic Institute (WPI) in 1996. The testbed connects five buildings with inter-LAN bridges using different technologies. Inside each building access points provide coverage to the laptops that are carried by the students. The professor broadcasts his/her image and writings on the electronic board to allow students to participate in the wireless classroom from different buildings on the campus. The entire wireless network is connected to the backbone through a router to isolate the traffic for traffic monitoring experimentations.

Today, horizontal markets for the WLAN industry mainly focus on WLAN as an alternative to LANs wherever the additional cost of the wireless solution is justifiable. This occurs for example in installations with frequent relocations where the additional cost of the WLAN solution is justified by the relocation costs of the wired solution. Temporary networking such as registration sites at conferences or fairs (jobs, food, etc.) is another example where the wireless solution is preferred to the less expensive and more reliable wired alternative. Buildings with difficult or impossible-to-wire situations, such as marble buildings or historical monuments where drilling for wiring is not favored, provides another example where WLAN is justifiable. The most prominent incentive for WLANs in vertical markets is the general use of laptops at home and in the offices.

2.4. New Interest from Military and Service Providers

In the mid-1990s, when the WLAN industry was struggling to find a market, a new wave of interest for WLAN came from the U.S. Department of Defense for



Figure 2. The experimental NSF sponsored WCAN at WPI.

military applications and from the European Community for commercial applications. These projects poured a considerable amount of research investments that further brightened the future of this industry [10].

The incentive for the military was to discover new horizons for implementation of global mobile military networks that support integration of computing and positioning systems. For instance, the InfoPAD project at the University of California, Berkeley [11] was one of the early WLAN DARPA projects. The environment is like a battleship equipped with a number of computing facilities. Soldiers in the environment are carrying InfoPADs that are small asymmetric communication devices carrying user instructions to the computing backbone to initiate computational operations whose results are downloaded to the PAD. A main challenge in this project was the implementation of reasonable size PADs capable of supporting multimedia applications. BodyLAN [12] was another DARPA-sponsored WLAN project initiated at BBN, Cambridge, Massachusetts. This project intended to design a low-power network capable of monitoring vital human body condition information (heartbeat, temperature, etc.) and communicating this information to other soldiers in the proximity. A more recent DARPA project was the Small Unit Operations/Situation Awareness Systems (SUO/SAS). The goal was to design an integrated telecommunication and geolocation network for modern fighting scenarios. The technical

challenges included providing accurate indoor position information [13] and communicating situation awareness information to the war fighters. This system was expected to provide a full communication and positioning link to the soldier operating inside a building.

The commercial interest from the European Community (EC) was initiated by the equipment manufacturers seeking solutions for the service providers that were keen on incorporating higher data rate services into the evolving rich cellular industry. In the mid-1990s, both commercial service providers and military network designers believed that the future of backbone networks would be an end-to-end ATM based network. In response to this perception, wideband local networking industries initiated the wireless ATM movement.

From the application point of view, service providers intend to integrate WLAN products into their existing services. A popular scenario used in HIPERLAN-2 to represent the service providers point of view is shown in Fig. 3. In this scenario it is assumed that a WLAN user carries his/her laptop in the office, home, and in public places (airports, train stations, etc.). In the home and office, the laptop connects to a free network whose infrastructure is owned by the user or his/her company. In public places, such as an airport or other transit buildings, WLAN access points belonging to a service provider can provide high-speed access. A widearea backbone wireless network could also provide the



Figure 3. The service provider's view of LANs.

connection but at lower data rates. In all public places the service provider that owns the infrastructure will charge the user. One of the technical challenges for the implementation of this scenario is the vertical roaming among different networks [14]. Another challenge is to incorporate roaming and tariff mechanisms for WLANs. These issues are currently under investigations.

2.5. A New Explosion of Market and Technology

During 1998 and 1999, interest in WLANs exploded. The WLAN industry that relied almost exclusively on a North American market with a market only a fraction of that of the cellular industry, suddenly attracted widespread attention from Japan and the EC and a renewed interest in the United States. There is now growing hope for a WLAN market comparable to that of the cellular industry. In Japan, small office spaces promoted the popular usage of laptops to replace PCs. The natural networking solution for laptops was nothing except WLANs. In the EC, the rich cellular industry started considering WLANs as part of their next generation high-speed packet data services. The interest is twofold. WLANs provide a practical higher-speed solution and operate in unlicensed bands free of charge while the cost of licensed bands is constantly increasing. In the North American continent, the successful growth of broadband Internet access to the homes opened a new window for a sizable market in home networking. This perception trend for a new sizable market was strengthened by the emergence of new low-power personal-area ad hoc wireless networking technologies such as Bluetooth and ultrawideband (UWB) for local distribution, LMDS for home access, and indoor positioning for a variety of applications. The availability of low-power, low-cost wireless chip sets started a new revolution in consumer product development, raising hope for sales in the order of hundreds of millions of chip sets per year. All together these hopes initiated a Gold Rush in chip manufacturing for WLAN and WPAN applications that is still ongoing. Technical directions for this industry remain as providing for higher data rates, more comprehensive coverage, less interference, and lower cost.

2.6. WLAN Standards

Wireless LANs provide very high data rates (≥ 1 Mbps) in local areas (<100 m) for access to wired LANs and high speed Internet. Today, all successful wireless LANs operate in unlicensed bands that are free of charge and rigorous regulations. Sine the late 1990s, considering that PCS bands were auctioned at very high prices, wireless LANs have attracted renewed attention.

Table 2 provides a summarizes the IEEE 802.11 and HIPERLAN standards for wireless LANs. IEEE standards include 802.11 and 802.11b operating at 2.4 GHz and 802.11a operating at 5 GHz. Both HIPERLAN-1 and -2, developed under ETSI, operate at 5 GHz. The 2.4-GHz products use spread-spectrum technology to support data rates ranging from 1 to 11 Mbps. HIPERLAN-1 uses GMSK modulation with DFE signal processing at the receiver and supports up to 23.5 Mbps. IEEE 802.11a and HIPERLAN-2 use an OFDM physical layer to support up to 54 Mbps. The access method for all 802.11 standards is the same and includes CSMA/CA, point coordination function (PCF), and request to send (RTS)/clear to send (CTS).

The access method for HIPERLAN-1 is comparable to 802.11, but the access method for HIPERLAN-2 is a voice-oriented access technique suitable for integration of voice and data services. IEEE 802.11, IEEE 802.11b, and HIPERLAN-1 are completed standards, and IEEE 802.11 and 11.b are today's dominant products in the market. IEEE 802.11a and HIPERLAN-2 are still under development. The IEEE 802.11 and HIPERLAN standards can be considered as the second generation of wireless LANs, while OFDM wireless LANs are forming the next generation of these products.

3. WPANs

3.1. Introduction

The first announced personal-area network (PAN) was the BodyLAN that emerged from a DARPA project in the mid-1990s. It was specified as a low-power, smallsize, inexpensive, modest-bandwidth solution that could connect personal devices in many collocated systems within a range of ~ 5 f [12]. Motivated by the BodyLAN project, a WPAN group originally started in June 1997 as a part of the IEEE 802.11 standardization activities. In January 1998, the WPAN group published the original functionality requirements. In May 1998, the study group invited participation from several related groups such as WATM, Bluetooth, HomeRF, BRAN (HIPERLAN), IrDA (IR short-range access), IETF (Internet standardization), and WLANA (a marketing alliance of WLAN companies in the United States). Only the HomeRF and Bluetooth groups responded to the invitation. In March 1998, the Home RF group was formed. In May 1998, the Bluetooth development was announced and a Bluetooth special group was formed within the WPAN group [15]. In March 1999, the IEEE 802.15 was approved as a separate group in the 802 community to handle WPAN standardization. At the time of this writing, IEEE 802.15 WPAN has four subcommittees on Bluetooth, coexistence, high data rate, and low data rate.

3.2. What Is IEEE 802.15 WPAN?

The 802.15 WPAN group is focused on development of short-distance wireless networks used for networking of portable and mobile computing devices such as PCs, personal digital assistants (PDA), cellular phones, printers, speakers, microphones and other consumer electronics. The WPAN group intends to publish standards that allow these devices to coexist and interoperate with one another and other wireless and wired networks in an internationally acceptable frequency band of operation.

The original functional requirement published in January 22, 1998 was based on the BodyLAN project and specified devices with [15]:

- Power management: small current consumption
- Range: 0-10 m
- Speed: 19.2–100 kbps (actual)
- Small size: ~ 0.5 in.³, no antenna
- Low cost: relative to target device
- Should allow overlap of multiple networks in the same area
- Networking support for a minimum of 16 devices

These specifications well fit the Bluetooth specifications, a technology announced after this premier announcement. The initial activities in the WPAN group included both the HomeRF and Bluetooth groups, but today HomeRF maintains its own Website at *www.homerf.org*. IEEE 802.15 WPAN includes four taskgroups. The first taskgroup is based on Bluetooth and aims to define PHY and MAC specifications for wireless connectivity among fixed, portable, and moving devices within or entering a *personal operating space* (POS), the space about a person or object that typically extends up to 10 m in all directions and envelops the person whether stationary or in motion. This taskgroup will address quality of service to support a variety of traffic classes.

The second taskgroup focuses on coexistence between WPAN and 802.11 WLANs. This group is developing a coexistence model to quantify the mutual interference and a coexistence mechanism to facilitate coexistence between an IEEE 802.11 WLAN and an IEEE 802.15 WPAN device. One goal of the WPAN group is to achieve a sufficient level of interoperability between a WPAN and an 802.11 device to allow transfer of data between the two devices.

The third taskgroup works on PHY and MAC layer specifications for high rate (HR) WPANs with data rates higher than 20 Mbps. This standard will provide for low-power, low-cost solutions that address the needs of portable consumer digital imaging and multimedia applications. This standard aims at providing compatibility with the Bluetooth specification of the taskgroup one. This standard is expected to be completed by early 2002.

The fourth taskgroup is chartered to investigate ultralow complexity, ultra-low-power consumption, ultra-lowcost PHY/MAC-layer specification to support data rates of up to 200 kbps. Potential applications are sensors, interactive toys, smart badges, remote controls, and home automation. This taskgroup may also address location tracking capabilities required to support the use of smart tags and badges.

3.3. What Is Home RF?

According to the standard committee, the mission of the HomeRF working group is to provide the foundation for a broad range of interoperable consumer devices by establishing an open industry specification for wireless digital communications between PCs and consumer electronic devices anywhere in and around the home.

Figure 4 represents the overall vision of HomeRF. The architecture can support both ad hoc and connected networks. In a popular home setup, the Internet access and PSTN connection arrives at a control HomeRF distribution box that can support HomeRF wireless as well as HPNA networks. The HomeRF wireless network can accommodate isochronous clients interconnecting up to six cordless telephone devices and asynchronous clients interconnecting a number of data devices. The two major competitors for HomeRF are HIPERLAN-2 and Bluetooth. As compared with HIPERLAN-2, the HomeRF solution provides smaller data rates (≤ 2 Mbps as opposed to 54 Mbps in HIPERLAN-2) that cannot support wireless transmission of video for TV and VCR applications. Compared with Bluetooth, HomeRF provides higher data rates, but Bluetooth was introduced as an inexpensive chip set that early on attracted a large alliance.

The HomeRF workgroup has developed a specification for wireless communications in the home called *shared wireless access protocol* (SWAP). The SWAP specification defines a new common interface to support wireless voice and data networking in the home. The SWAP specification is an extension of DECT (TDMA) for voice, and a relaxed 802.11 (CSMA/CA) for high-speed data applications.

The reader interested in more details on HomeRF is referred to Ref. 16.

3.4. What Is Bluetooth?

Bluetooth is an open specification for short-range wireless voice and data communications that was originally developed for wire replacement in personal area networking to operate all over the world. In 1994, the initial study for development of this technology started at Ericsson, Sweden. In 1998, Ericsson, Nokia, IBM, Toshiba, and Intel formed a special-interest group (SIG) to expand the concept and develop a standard under IEEE 802.15 WPAN. In 1999, the first specification, v1.0b, was released and then accepted as the IEEE 802.15 WPAN standard for a 1-Mbps network. At the time of this writing over 1000 companies participate as members in the Bluetooth SIG, and a number of companies all over the world are developing Bluetooth chip sets. Marketing forecasts indicate penetration of Bluetooth in more than 100 million cellular phones and several million of other communication devices. The IEEE 802.15 is also studying coexistence and interference between Bluetooth and IEEE 802.11 products operating at 2.4 GHz.

The story of the origin of the name Bluetooth is interesting and worth mentioning. "Bluetooth" was the nickname of Harald Blaatand (940–981 A.D.), King of



peer-peer devices

Figure 4. Overview of the HomeRF vision.

Denmark and Norway. When Bluetooth was introduced to the public, a stone carving erected from Harald Blaatand's capital city Jelling was presented [17]. This strange carving was interpreted as Bluetooth connecting a cellular phone and a wireless notepad in his hands. This picture was used to symbolize the vision of using "Bluetooth" to connect personal computing and communication devices. Bluetooth, the king, was also known as a peacemaker and a person who brought Christianity to Scandinavians to harmonize their beliefs with the rest of Europe. This fact is used to symbolize the need for harmony among manufacturers of WPANs around the world to support the growth of the WPAN industry.

Bluetooth is the first popular technology for short-range ad hoc networking designed for integrated voice and data applications. As compared with WLANs, Bluetooth has a lower data rate but has an embedded mechanism to support voice applications. As compared with 3G cellular systems, Bluetooth is an inexpensive personal-area ad hoc network operating in unlicensed bands and owned by the user.

The Bluetooth SIG considers three basic application scenarios [17]. The first scenario is the wire replacement to connect a personal computer or laptop to the keyboard, mouse, microphone, and notepad. As the name indicates this avoids the problem of multiple short-range wirings surrounding today's personal computing devices. The second scenario is an ad hoc network of several different users in a very short range of each other such as in a conference room. WLAN standards and products are also commonly considered for this scenario. The third scenario considers Bluetooth access points to redistribute widearea voice and data services provided by cellular networks, wired connections or satellite links, in a fashion similar to that of the WLAN scenario in an airport. Contrary to 802.11 however, Bluetooth has provisions for both voice and data and can be used as an integrated voice/data access point to connect to both voice and data backbone infrastructures. The HIPERLAN-2 standard will provide a more expensive version of similar connections that supports a larger number of users and higher data rates.

The topology of a Bluetooth network is referred to as *scattered ad hoc topology*. In a scattered ad hoc environment a number of small networks, each supporting a few terminals, can coexist and possibly interoperate with each another. Bluetooth specifications have been selected to operate in the unlicensed ISM bands at 2.4 GHz. The advantage is the worldwide availability of the bands. The disadvantage is the existence of other users, in particular IEEE 802.11 and 802.11b products in the same band. At the time of this writing, a subcommittee of the IEEE 802.15 is working on the interference issues related to Bluetooth and IEEE 802.11 and 802.11b.

4. HOME NETWORKING

In a house, the number of devices that are connected together or need to communicate with the outside world presents a new challenge. Figure 5 provides an illustration of how diverse and fragmented networking connections can become at home. The house is connected to the



Figure 5. Today's fragmented home access and distribution networks.

PSTN for telephone services, to the Internet for Web access, and to cable network for multichannel TV services. Inside the home, computers and printers are connected to the Internet through voiceband modems, xDSL services, or cable modems. The telephone services and security systems are connected through the phone line. The TV is connected to the multichannel services through HFC cables or satellite dishes. Audio and video entertainment equipment such as a videocamera and a stereo system, and computing systems such as laptops are either isolated or have proprietary wired connections.

This fragmented networking environment has prompted a number of initiatives to create a home network. The home networking industry started in the late 1990s by the design of the so-called home gateways to connect the increasing number of computer appliances, and distribute a single Internet connection. The number of home networks in the United States is expected to almost double each year. This industry has two distinct branches: home access and home distribution segments. The home access technology employs different wireless and wired alternatives to secure a broadband Internet access to the home gateway to be distributed to the users information appliances.

The home distribution or *home-area network* (HAN) interconnects all the home appliances and connects them to the Internet through the home gateway. For the access industry, it is expected that 80% of U.S. households will have a broadband data access by the year 2004. For the distribution industry it is expected that the number of sold "information appliances" will exceed the

sold number of PCs by the year 2002. It is also expected that to interconnect PCs and information appliances to the broadband services, 10 million home networks will be installed by the year 2004.

4.1. What Is a HAN?

The home-area network (HAN) provides an infrastructure to interconnect a variety of home appliances and connect them to the Internet through a central home gateway. A number of home appliances are emerging in the market that are in need of a HAN. Figure 6 provides an overview of these appliances classified into logical groups.

Home computing equipment is used for computing and Internet transaction interface access and includes PCs, laptops, printers, scanners, and QuickCAMs. If there is no home distribution network all the equipment is connected together either through the PC or laptop ports. A home computing network allows multiple computers as well as multiple devices to connect with a network protocol. A wireless network allows flexibility in installation and relocation of these devices in different rooms of a home. Phone appliances used for two-way conversations are cordless telephones, intercommunication devices, and standard wired telephone sets. All the telephone services have an interface to communicate with the PSTN. Currently they are connected through the home telephone wiring to the PSTN. With a HAN these devices can share the home access medium (cable or TP) allowing one service provider to bring both data and voice services. The entertainment audiovisual appliances include TVs, stereo systems, CD players, VCRs, DVD players, tape



Figure 6. Classification of home equipment demanding networked operation.

recorders, camcorders, speakers, and headphones. These devices communicate through their own protocols such as the more recent IEEE 1394 or HAVi [18]. The security system includes motion detectors, door pins, system control panel, camera, and alarm that are networked separately using protocols such as EIA-600 CEBus [19] or the industrial initiative EIA-709 LonWorks [20]. Currently, the access to these systems is through the telephone lines that initiate emergency request alarms to the police stations. Appliance manufacturers are working on "smart" appliances capable of communication. This intelligence allows remote checking test for maintenance (e.g., receiving alarm that refrigerator's filter needs to be replaced) and remote control of operation (e.g., turning a washing machine on from outside). Besides LonWorks, another industrial initiative comes from Merloni's Ariston Digital and is called Web-ready appliances protocol (WRAP) [21]. Another wave of interest has been initiated by utility companies for distance utility metering. The electricity, gas, water, or fuel companies would like to read the meter at home for billing or other needs (e.g., refilling the gas tank). One of the solutions is to communicate this information through the HAN and its access to the Internet. More recently, a number of startup companies have been designing indoor locating systems that can be used for distance monitoring of children, the elderly, and pets or for navigating the blind. These systems are expected to be integrated with the computing networks to provide access through the Internet.

4.2. Why Do We Need a HAN?

The existing LANs designed for office environments do not provide a good solution for home networking. The application diversity, network requirements, building infrastructure, and market size of HANs are distinctly different from that of LANs. At home, the number of users of the network is much smaller than in the offices but the diversity of the device types and their bandwidth requirements are much larger than in the offices. The diversity of bandwidth requirement can range from multichannel video to monthly meter reading. In an office, computing devices are predominant, but the home environment includes new applications that were not needed for LANs such as audiovideo broadcasting and positioning/navigation. Office environments are larger than residential homes and are made up of material more concrete than that found in the home. Therefore physical wiring and wireless coverage in the homes is easier than in offices. Homeowners are more reluctant to allow service workers to enter their homes, and they cannot afford a network manager to operate their network. The number of homes is an order of magnitude larger than that of the offices, so the market for home networking is expected to be much larger than for LANs.

These specific requirements on home applications impose certain constraints on the design of HANs. A HAN needs to be *user-friendly* because it is used and managed by nonprofessionals with limited technical skills and small budget size. A HAN must be low cost, easy to install and relocate, and easy to upgrade. In terms of *performance* a HAN should enable multimedia applications and be capable of accommodating legacy voice and data services. A HAN also needs to be *flexible and scalable* to allow location independent easy-to-reconfigure networks without significant performance degradations. To avoid eavesdropping a HAN also needs *security and privacy* provisions.

4.3. HAN Technologies

For the offices, a company recognizes the need for networking, decides on installing a network, opens a budget for expensive wiring, and installs the LAN infrastructure. In the homes, users gradually build their networks at their leisure with an investment that is spread over a relatively long period. An average customer of home network does not spend a sizable budget on the wiring to develop an infrastructure. Therefore, the trend in today HANs is to avoid adding any new wiring by either using existing wires or by using a wireless solution. The existing wirings at homes are twisted-pair (TP) telephone wirings, power-line wirings and cable TV wirings. The phone-line wirings have a relatively good distribution and in most modern homes at least one telephone outlet can be located in every room. The wiring for phone lines is voice grade TP that is suitable for Ethernet connections. However, this line is used for carrying regular phone and xDSL services that will interfere with an Ethernet signal. Power-line wirings are even better distributed because every room has several power outlets. However, the quality of the line is poorer and the level of noise is much larger than in TP phone-line wirings. The characteristics of the lines impose limitations on data rates that can be overcome only by using more complex transmission techniques. Existing cable TV wirings have very restricted distribution and only a few outlets are available in each household. This wiring is used for multichannel TV distribution that will interfere with an Ethernet signal. The expensive broadband cable TV modems can be used to overcome this problem. Because of its limited distribution and expensive modem requirements, cable TV wirings are not considered seriously for home distribution. Wireless solutions appear ideal for home networking. The ease of installation and relocation provides an excellent solution. Challenges for wireless is reliability, bandwidth, coverage, and interference. Comparing wired and wireless solutions, current wired HANs can be implemented using less expensive network cards and are expected to support higher data rates. Wireless HANs provide ideal ad hoc solutions that support portability.

4.3.1. HPNA. In outdoor areas the PSTN network has two parts, the TP analog access wirings and the backbone digital wiring connecting PSTN switches together. The digital segment of the PSTN fully utilizes the wiring while the traditional access wiring was using only \sim 4 kHz for analog POTS and the rest was unexplored until xDSL and HPNA were introduced. Further, today computers purchased with network capability as well as PCMCIA network cards for laptops are exclusively Ethernet-based. However, as we mentioned before, the TP phone line wirings at home are also used for analog voice and xDSL access. HPNA is an Ethernet-compatible LAN over the random-tree home phone lines. It uses a standalone adapter to connect directly to the in-home telephone jacks any device having an Ethernet 10base-T interface card, operating at 10 Mbps over the TP lines.

HPNA shares the TP line medium with POTS and xDSL using FDM. POTS uses the 20–3400 Hz band for analog voice transmission, xDSL uses a 25–1100 kHz to provide high-speed Internet access, and HPNA uses a 2–30 MHz band for home distribution networking. The HPNA is based on a patented physical-layer design that is more immune to high noise conditions in the home telephone wirings. The MAC layer for the HPNA is the same as the MAC layer of the IEEE 802.3 Ethernet. From the user's point of view, the HPNA network accommodates the exiting legacy Ethernet software and hardware. Only an adaptor is placed between the Ethernet connection and the phone plugs. The next step for HPNA is to boost the data rate to accommodate video applications.

4.3.2. Power-Line Modems. As compared with telephone wirings, power line wirings have the best wiring distribution in the home due to the superior number of electric outlets that can be found. The entire power line wiring is used for transmission of a 50-60-Hz waveform, and the rest is available for other applications. For many years power lines were used for low-data-rate (<100 kbps) control and security networks operating below 500 kHz. These systems were mostly using X-10 and the CEBus/CAL standards [17]. More recently, power lines are being considered for high data rate communications (>1 Mbps) and to operate above 1 MHz to provide adequate speeds for computer networking. This area is still in the preliminary stages of development with no clear standard initiative. European regulations prohibit power line signaling above 150 kHz due to potential interference with low-frequency licensed radio services. Current U.S. and Japanese regulations allow the use of a somewhat broader spectrum up to \sim 525 kHz where AM radios begin. In the power line, a low-frequency band of up to a few kilohertz is used for low-data-rate applications such as security, and a high-frequency band (1-30 MHz) is used for high-speed data communications.

Power lines suffer from tremendous interference from electrical appliances, high attenuation, reflection caused from varying input impedance, and multipath phenomena that makes communication over this medium as challenging as communication over radio channels. As a result, a variety of complex transmission techniques and medium-access protocols has been examined for different power line applications. Traditional FSK and QPSK are used in lower bands and more complex spread spectrum and OFDM modems are used in the higher bands [22]. The difficulty of the medium has complicated the development of low-cost solutions and is a drawback in growth of this market. More recently, smart appliances have been emerging in the market that have some builtin intelligence, can sense other appliances on the power lines, and can be accessed through the Internet. Electric companies are investigating using the outside AC lines to deliver various services such as meter reading, energy management, and even Internet access to the homes. The main current research thrust is to enable access to the control and security systems through the Internet.

4.3.3. Wireless Solution Alternatives. Compared with wired networks, wireless solutions can provide mobility

and coverage to the home as well as the yard. Cordless telephones were very successful from the early days they appeared in the market. Since cordless telephones provide mobility and extended coverage, users have paid higher prices to purchase cordless telephones instead of wired phones. The other advantages are that wireless solutions are easy to install, relocate, scale, and maintain.

The introduction of wireless solutions to home security systems resulted in a sizable growth of that industry. From the user's point of view, wireless security systems are installed very quickly without additional holes in the walls or wires distributed in the home. Furthermore, selecting the location for placing a wireless product is more flexible and allows a better blending with the decoration of the home. Other advantages are that wireless solutions are easier to be expanded, moved to new homes, maintained, and upgraded.

The examples above described are selected applications where the wireless solution is preferred. However, there are numerous home applications and a wireless solution may not be the best one for all of them. Home automation network applications such as switching a light through remote control is done by sending a single bit (ON/OFF) through the power lines to an inexpensive ON/OFF X-10 switch attached to the lamp. Simple infrared remote control can provide the additional mobility by sending the control command to a general control box connected to the power lines. It is not difficult to find other examples involving the power lines or phone lines where wireless might not be beneficial. The important conclusion from this discussion is that in home networking we want to avoid new wiring, not avoid using the existing wires. The home therefore can be seen as a nonhomogeneous environment for networking.

The principal candidates for wireless home networking are WLANs and WPANs discussed in the previous sections. There are a number of home-specific challenges for wireless home networking. The use of noncompatible wireless devices operating in the same unlicensed band can become an issue if they interfere with each other. Handling interference in such cases is an important issue. Another issue is that as we move to higher frequencies of operation in the 5-GHz range to support higher data rates, the coverage of the wireless solution may become a challenge. Wireless home networking needs designing inexpensive reconfigurable devices and internetworking between diverse mediums using different protocols. The network to incorporate cable TV applications requires high transmission rates as well as new delivery techniques to the TV set. Today, the coaxial cable that connects to the converter box on the TV set carries around a hundred analog video channels. This extent of bandwidth is not feasible in a wireless medium, and therefore, the entire system needs to be redesigned.

4.4. Home Access Networks

The early home access technology was based on voice band modems over the phone line. Today, broadband home access with data rates on the order of 10 Mbps is provided through cable modems or xDSL services. The cable distribution in the residential areas has a bus topology that is optimally designed for one-way TV signal distribution. The bus carries all the stations in the neighborhood. Cable modems use a bandpass channel allocated to a TV channel to provide high data rates for transmission using QAM modulation. Broadband cable services use one of the video channels and a reverse channel to establish a two-way communication and access to Internet. The xDSL service uses a 25–1100-kHz band on the phone line and multisymbol QAM modulation to support high data rates to the users. The topology of the telephone line is a star topology that connects every user directly to the end office, where the xDSL data is directed to Internet through a router.

Higher-speed wireless home access uses a local multipoint distribution system (LMDS) or even existing WLAN inter-LAN bridges to provide the service. The advantage of using a fixed wireless solution is that it does not involve wiring under the streets. The wireless solution is certainly attractive when there is no existing wiring in the neighborhood or when obtaining city permission might be troublesome. The HIPER-ACCESS preprogram in the EC and IEEE 802.16 are currently studying the specifications for the next generation of wireless local access. Other wireless alternatives are direct satellite TV broadcasting and 3G wireless networks. Direct broadcast suffers by the lack of a reverse channel and high delays that challenges the implementation of broadband services. The high-speed 3G wireless packet data services are expected to provide up to 2 Mbps, suitable for Internet access. However, besides a lower data rate than the others, these networks will be using licensed bands, and ultimately may be expensive as well. Figure 7 summarizes the existing solutions for the home access technologies.

5. INDOOR GEOLOCATION

5.1. Introduction

An important evolving technology recent has been indoor geolocation technology for both military and commercial applications. There is an increasing need in hospitals to locate patients or expensive equipment and in homes



Figure 7. Broadband home access alternatives.
to locate children and equipment. In the Department of Defense, Small Unit Operation/Situation Awareness Systems require a modern warfighter to be able to communicate in an urban environment, and his/her position to be known at all times including when he is inside a building. An indoor geolocation system can also be crucial for the safety of a firefighter entering a blazing building when his/her position is being monitored. Similar scenarios can be found in other public safety agencies. These incentives have lead to research in indoor geolocation systems [13,23,24]. In an indoor environment, traditional GPS or E-911 location systems do not work properly due to severe multipath effects. As a result, dedicated indoor systems have to be developed to provide accurate indoor geolocation services.

5.2. Wireless Geolocation Methods and Metrics

Most geolocation system architectures and methods developed for cellular systems are applicable for indoor geolocation systems although special considerations are needed for indoor radio channels. The most widely used wireless geolocation metrics include angle of arrival (AoA), time of arrival (ToA), time differences of arrival (TDoA), received-signal strength (RSS), and received-signal phase (RSP). In the following, we present an overview of overall system architectures and basic concepts of geolocation metrics as well as corresponding geolocation methods.

5.2.1. Overall System Architecture. Similar to cellular geolocation systems, the architecture of indoor geolocation systems can be roughly grouped into two main categories: mobile-based architecture and network-based architecture. Most of indoor geolocation applications proposed to date have been focused on a network-based system architecture as shown in Fig. 8 [25,26].

The geolocation base stations (GBSs) extract location metrics from the radio signals transmitted by the mobile station and relay the information to a geolocation control station (GCS). The connection between GBS and GCS can be either wired or wireless. Then the position of the mobile station is estimated, displayed and tracked at the GCS. With the mobile-based system architecture, the mobile station estimates self-position by measuring the received radio signals from multiple fixed GBS. Compared to a mobile-based architecture, the network-based system has the advantage that the mobile station can be implemented as a simple-structured transceiver with small size and low power consumption, easily carried by people or attached to valuable equipments as a tag.

5.2.2. Angle of Arrival. The AoA geolocation method uses simple triangulation to locate the transmitter as shown in Fig. 9.

The receiver measures the direction of the received signals (i.e., angle of arrival) from the target transmitter using directional antennas or antenna arrays. If the accuracy of the direction measurement is $\pm \theta_s$, AoA measurement at the receiver will restrict the transmitter position around the line-of-sight (LoS) signal path with an angular spread of $2\theta_s$. AoA measurements at two receivers will provide a position fix as illustrated in Fig. 9. We can clearly observe that given the accuracy of AoA measurements, the accuracy of the position estimation depends on the transmitter position with respect to the receivers. When the transmitter lies between the two receivers, AoA measurements will not be able to provide a position fix. As a result, more than two receivers are normally needed to improve the location accuracy. For a macrocellular environment where the primary scatters are located around the transmitter and far away from the receivers, the AoA method can provide acceptable location accuracy [27]. But dramatically large location errors will occur if the LoS signal path is blocked and the AoA of a reflected or a scattered signal component is used for



Figure 9. Angle-of-arrival geolocation method.



Figure 8. Overall architecture of indoor geolocation system.

estimation. In indoor environments, the LoS signal path is usually blocked by surrounding objects or walls. Thus the AoA method alone cannot provide sufficient accuracy for indoor geolocation systems.

5.2.3. Time of Arrival and Time Difference of Arrival. The ToA method is based on estimating the propagation time of the signals from a transmitter to multiple receivers. Several different methods can be used to obtain ToA or TDoA estimates, including pulse ranging [28,29], phase ranging [28], and spread-spectrum techniques [25,30].

Once the ToA is measured, the distance between the transmitter and receiver can be determined simply since the propagation speed of the radio signal is approximately the speed of light. The estimated distance at the receiver will geometrically define a circle, centered at the receiver, of possible transmitter positions. ToA measurements at three receivers will provide a position fix and given receiver coordinates and distances from the transmitter to receivers, the transmitter coordinates can be easily calculated. In indoor environments, the strongest signal received can come from a longer reflection path and not from the direct line-of-sight path. The ToA-based distance estimates are therefore always larger than the true distance between the transmitter and the receiver as illustrated in Fig. 10, where \hat{r}_1, \hat{r}_2 , and \hat{r}_3 are the estimated distances and r_1 , r_2 , and r_3 are the true distances. Three ToA measurements determine a region of possible transmitter position as shown in Fig. 10. A nonlinear least-square (NLLS) method is usually used to obtain the best estimation iteratively [25,27]. A constrained NLLS algorithm is also available that makes use of the fact that ToA-based distance estimates are always larger than the true distance [31]. As in the AoA method, more than three ToA measurements are needed to improve the accuracy of position estimation.

Instead of using ToA measurements, time difference measurements can also be employed to locate the receiver position. A constant time difference of arrival (TDoA) for two receivers defines a hyperbola, with foci at the receivers, on which the transmitter must be located. Three or more TDoA measurements provide a position fix at the intersection of hyperbolas. NLLS method can also be used to obtain the best estimation of the transmitter position.



Figure 10. Time of Arrival geolocation method.

Some other methods have been used to solve the hyperbolic position estimation problem [32–35]. Compared to the ToA method, the main advantage of the TDoA method is that it does not require the knowledge of the transmit time from the transmitter while the ToA method requires it. As a result, strict time synchronization between the transmitter and receivers is not required. However, the TDoA method requires time synchronization among all the receivers.

5.2.4. Received-Signal Strength. If the power transmitted by the mobile terminal is known, measuring the received-signal strength (RSS) at the receiver can provide an estimate of the distance when the path-loss characteristics of the environment are known. As in the ToA method, the measured distance will determine a circle, centered at the receiver, on which the mobile transmitter must lie. Three RSS measurements will provide a position fix for the mobile. As a result of shadow fading effects, the RSS method results in large range estimation errors. The accuracy of this method can be improved by utilizing a premeasured received signal strength contour centered at the receiver [36]. A fuzzy-logic algorithm has been shown [37] to be able to significantly improve the location accuracy.

5.2.5. Received-Signal Phase. Signal phase is another possible geolocation metric. It is well known that with the aid of reference receivers to measure the carrier phase, differential GPS (DGPS) can improve the location accuracy from ~ 20 m to within 1 m compared to the standard GPS, which only uses pseudorange measurements. One problem associated with the phase measurements lies in the ambiguity resulting from the periodicity of the signal phase while the standard pseudorange measurements are unambiguous. Consequently, in the DGPS, the ambiguous carrier phase measurement is used to fine-tune the pseudorange measurement. A complementary Kalman filter is used to combine the low-noise ambiguous carrier phase measurements with the unambiguous but noisier pseudorange measurements [38]. For indoor geolocation systems, it is possible to use the received-signal phase method together with the ToA/TDoA or the RSS method to fine-tune the location estimate. However, unlike DGPS, where the LoS signal path is always observed, multipath and non-line-of-sight conditions in the indoor environment can cause large errors in phase measurements.

BIOGRAPHIES

Kaveh Pahlavan, is a professor of ECE and CS, and director of the CWINS laboratory at WPI, Worcester, Massachusetts. He is also a visiting professor of TLab and CWC, University of Oulu, Finland. He is the principal author of the *Wireless Information Networks*, John Wiley and Sons, 1995 (with A. Levesque) and *Principles of Wireless Networks*—A Unified Approach, Prentice Hall, 2002 (with P. Krishnamurthy). He has published numerous papers, served as a consultant to a number companies, and sits in the board of a few companies. He is the editor in chief and founder of the *International Journal of Wireless Networks*, the founder of

the IEEE Workshop on Wireless LANs, and a cofounder of the IEEE PIMRC conference. For his contributions to evolution of the wireless networks he has received the Weston Hadden Professor of ECE at WPI, been elected as a fellow of the IEEE in 1996, become a fellow of Nokia in 1999, and received the first Fulbright–Nokia Scholarship Award in 2000. Because of his inspiring visionary publications and his international conference activities for the growth of the wireless LAN industry he is referred to as one of the founding fathers of the wireless LAN industry. Details of his contributions to this field are available at www.cwins.wpi.edu.

Jacques Beneat received his Ph.D. degree in electrical and computer engineering from Worcester Polytechnic Institute (WPI), Massachusetts, in 1993 with focus on the design of data communication circuits and advanced microwave structures for satellite communications. During several years, he served as adjunct faculty in the University of Bordeaux, France, Worcester State College, and WPI. In 1996, he joined the Center for Wireless Information Network Studies at WPI, and during the past five years, he has served as a research scientist with focus on indoor radio propagation measurements and modeling using ray tracing techniques, real-time channel simulators and performance evaluation for emerging wireless networks. He was the general conference secretary for the ninth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) in 1998, for the third IEEE Workshop on Wireless LANs in 2001, and was a guest editor of the International Journal of Wireless Information Networks for a special series on implementation issues in wireless communications in 1997-98. In 2002, he will be joining the EE Department at Norwich University in Vermont.

XINRONG LI (xinrong@wpi.edu) received his B.E. and M.E. from the University of Science and Technology of China, Hefei, China, and the National University of Singapore, Singapore, in 1995 and 1999, respectively. He is currently pursuing his Ph.D. degree in wireless communications and networks at Worcester Polytechnic Institute (WPI), Worcester, Massachusetts. He has been working as research assistant in the Center for Wireless Information Network Studies (CWINS), WPI, and his recent research has been focused on indoor geolocation, statistical signal processing, and wireless networks. During the summer of 2002, he was also involved in the development of 3G All-IP CDMA2000 1xEV-DO Wireless Data Networks at Airvana, Inc., Chelmsford, MA.

BIBLIOGRAPHY

- 1. F. R. Gfeller and U. H. Bapst, Wireless in-house data communication via diffuse infrared radiation, *Proc. IEEE* **67**(11): 1474-1486 (Nov. 1979).
- P. Ferert, Application of spread spectrum radio to wireless terminal communications, *Proc. NTC*, Houston, TX, Dec. 1980, pp. 244-248.
- K. Pahlavan, Wireless communications for office information networks, *IEEE Commun. Mag.* 23(6): 19–27 (June 1985).

- M. J. Marcus, Regulatory policy considerations for radio local area networks, *IEEE Commun. Mag.* 25(7): 95–99 (July 1987).
- 5. K. Pahlavan, Wireless Intra-office Networks, ACM Transactions Office Information Systems, July 1988.
- M. Kavehrad and P. J. McLane, Spread spectrum for indoor digital radio, *IEEE Commun. Mag.* 25(6): 32–40 (June 1987).
- 7. K. Pahlavan and A. Levesque, Wireless Data Communication, invited paper, IEEE Proc., Sept. 1994.
- 8. K. Pahlavan, T. Probert, and M. Chase, Trends in local wireless networks, invited paper, *IEEE Commun. Soc. Mag.* (March 1995).
- First IEEE Workshop on Wireless LANs, Worcester, Mass., May 1991, http://www.cwins.wpi.edu/wlans91/index.html.
- K. Pahlavan, A. Zahedi, and P. Krishnamurthy, Wideband local access: WLAN and WATM, *IEEE Commun. Mag.* (Special Series on Wireless ATM), (Nov. 1997).
- 11. S. Narayanaswamy et al., Application and network support for infopad, *IEEE Pers. Commun. Mag.* (March 1996).
- L. Dennison, Body LAN, a wearable personal network, Proc. 2nd IEEE Workshop on Wireless LANs, Worcester, MA, Oct. 1996.
- K. Pahlavan, P. Krishnamurthy, and J. Beneat, Wideband radio propagation modeling for indoor geolocation applications, *IEEE Commun. Mag.* 60–65 (April 1998).
- K. Pahlavan et al., Handoff in hybrid mobile data networks, IEEE PCS Mag. 7(2): 34-47 (April 2000).
- T. Siep, I. Gifford, R. Braley, and R. Heile, Paving the way for personal area network standards: An over view of the IEEE P802.15 Working Group for Wireless Personal Area Networks, *IEEE Commun. Soc. Mag.* (Feb. 2000).
- K. Negus, A. Stephens, and J. Lansfield, HomeRF: Wireless networking for the connected home, *IEEE Commun. Soc. Mag.* (Feb. 2000).
- 17. Bluetooth SIG, http://www.bluetooth.com.
- HAVi (home audiovideo interoperability), http://www.havi. org/home.html.
- 19. CEBus—Consumer Electronics Bus, http://www.cebus. org/.
- 20. LonWorks Technology Platform, http://www.echelon.com/.
- 21. WRAP—Ariston Digital Web-ready appliance protocol, http://www.margherita2000.com/wrap/uk/main/index. htm.
- 22. Intellon High Speed Power Line Communications, white paper, http://www.intellon.com/.
- P. Krishnamurthy, K. Pahlavan, and J. Beneat, Radio propagation modeling for indoor geolocation applications, *Proc. IEEE PIMRC'98*, Sept. 1998.
- J. Werb and C. Lanzl, Designing a positioning system for finding things and people indoors, *IEEE Spectrum* 35(9): (Sept. 1998).
- 25. PinPoint Local Positioning System, http://www.pinpointco. com/.
- 26. PalTrack Tracking Systems, http://www.sovtechcorp.com/.
- J. Caffery, Jr. and G. L. Stuber, Subscriber location in CDMA cellular networks, *IEEE Trans. Vehic. Technol.* 47(2): (May 1998).

2692 TROPOSPHERIC SCATTER COMMUNICATION

- G. Turin, W. Jewell, and T. Johnston, Simulation of urban vehicle-monitoring systems, *IEEE Trans. Vehic. Technol.* VT-21: 9-16 (Feb. 1972).
- H. Hashemi, Pulse ranging radiolocation technique and its application to channel assignment in digital cellular radio, *Proc. IEEE VTC'91*, 1991, pp. 675-680.
- P. Goud, A. Sesay, and M. Fattouche, A spread spectrum radiolocation technique and its application to cellular radio, *Proc. IEEE Pacific Rim Conf. Communications, Computer* and Signal Processing, 1991, pp. 661–664.
- G. Morley and W. Grover, Improved location estimation with pulse-ranging in presence of shadowing and multipath excessdelay effects, *Electron. Lett.* **31**: 1609–1610 (Aug. 1995).
- W. H. Foy, Position-location solutions by Taylor-series estimation, *IEEE Trans. Aerospace Electron. Syst.* AES-12: 187–194 (March 1976).
- D. J. Torrieri, Statistical theory of passive location system, *IEEE Trans. Aerospace Electric Syst.* AES-20(2): (March 1992).
- 34. J. S. Abel and J. O. Smith, A divide-and-conquer approach to least-squares estimation, *IEEE. Trans. Aerospace Electric* Syst. 26: 423-427 (March 1990).
- Y. T. Chan and K. C. Ho, A simple and efficient estimator for hyperbolic location, *IEEE Trans. Signal Process.* 42(8): 1905-1915 (Aug. 1994).
- W. Figel, N. Shepherd, and W. Trammell, Vehicle location by a signal attenuation method, *IEEE Trans. Vehic. Technol.* VT-18: 105-110 (Nov. 1969).
- H.-L. Song, Automatic vehicle location in cellular communications systems, *IEEE Trans. Vehic. Technol.* 43(4): 902-908 (Nov. 1994).
- E. D. Kaplan, Understanding GPS: Principles and Applications, Artech House, Boston, 1996.

TROPOSPHERIC SCATTER COMMUNICATION

PETER MONSEN P.M. Associates Stowe, Vermont

1. INTRODUCTION

Prior to the launching of the first active repeater satellite TELSTAR on July 10, 1962, long-distance communication was limited to cable systems and radio communication techniques that exploited characteristics of either the ionosphere or the troposphere. Radio systems offered advantages of greater network flexibility, the ability to transverse difficult terrain, and less susceptibility to connectivity loss due to either catastrophic failure or sabotage. In high-frequency (HF) radio systems, ionospheric reflections made communication possible up to thousands of kilometers but with a limited bandwidth that would permit transmission of only one or two telephone channels.

When the transmitting and receiving antennas of a radio link are reciprocally visible, the link is classified as line-of-sight (LoS) and the received signal is reduced as the square of the distance. Beyond the LoS zone which can extend up to about 50 km, the received signal power is more severely reduced with distance because now the radio waves must bend or diffract with the curvature of the earth. In 1937 VanDerPool and Bremmer [1] determined losses due to diffraction over a spherical surface. These losses increase exponentially with distance beyond the LoS zone, so communication dependent on diffraction is limited to short distances beyond the horizon.

The introduction of higher-power radars during World War II, however, led to radio interferences at beyond LoS distances that was considerably larger than predicted by diffraction theory [2]. Some of the contribution to stronger signals was attributed to a refractive index (the ratio of the speed of light in a vacuum relative to the medium of interest) that decreased with height, thus increasing the "effective earth radius." Measurement of interference after World War II from frequency modulation radio stations and television stations confirmed that additional factors were contributing to the enhanced signal levels.

The troposphere extends from the ground to about 10 km and includes virtually all weather phenomena. In the 1950s it was recognized that the turbulence in the troposphere due to variations of meteorological characteristics such as temperature and humidity, would produce inhomogeneities in the refractive index. As a result of these refractive index variations within the scattering common volume corresponding to the intersection of the transmit and receive antenna beams, some fraction of the radio energy would be scattered in the direction of the receive antenna [3,4]. Experimental links implemented between 1950 and 1955 established that although the received signal was continuously variable it was permanently present so that reliable communication would be possible. Because of the scattered radio signals were weak, it was necessary to use very large antennas, sensitive low-noise receivers, and powerful klystron amplifiers with kilowatt outputs. The first "troposcatter" military and commercial links were installed from 1953 onward and provided up to 60 telephone channels in the 400-900-MHz frequency band over distances up to 300 km [5].

2. TROPOSPHERIC SCATTER RADIO LINKS

The provision of multiple voice channel circuits over long distances was the primary purpose of early troposcatter radio links. These links would be connected together to provide telephone conversations, for example, between the United States and Europe via a network of military links that traversed Greenland, Iceland, and Scotland and into Europe as far as Turkey.

In these troposcatter radio links the strongest signal is received when the antennas are aimed approximately at the horizon. The angle between the transmit and receive antenna beam centerlines at their intersection is called the scattering angle θ . In the LoS zone, for example, where the antennas are reciprocally visible this angle is zero. The transmission loss in troposcatter mode is proportional to $\theta^{-\alpha}$, where various theories predict α to be between $\frac{11}{3}$ and 5. Thus the antennas are pointed as low to the horizon as possible without suffering undo blockage. The antennas at both the transmitter and receiver were relatively large in order to successfully capture the weak scatter signal and provide reliable communication. On some of the longest links these antennas could be 120 ft in diameter.

The standard telephone channel has a 4-kHz bandwidth. In the earlier analog troposcatter systems these channels would be frequency-division multiplexed (FDM) and frequency modulated (FM) to a radiofrequency (RF) carrier between a few hundred megahertz and 5000 MHz. The power amplifier was typically a klystron power tube capable of generating kilowatts of power.

Digital systems used a time-division multiplex (TDM) of digitally converted voice channels. Modulation techniques such as quadrature phase shift keying (QPSK) could be used to convert the composite high-rate digital stream into an RF signal for subsequent amplification and transmission.

At the receiver, a low-noise amplifier was employed for each antenna input in order to increase the margin between the received signal and the ever-present thermal noise. Demodulation in analog systems of the received signal was accomplished in a frequency modulation discriminator that produced a voltage proportional to the instantaneous frequency. In a digital system, adaptive processing was included in the QPSK demodulation process in order to cope with time-delayed scattered components. After demodulation the composite detected signal would be demultiplexed and converted to a series of telephone channels. In an analog system the noise introduced by transmission impairments would be passed on to the next link and would accumulate over a series of tandem links. A digital system offered the advantage of regenerating (although with occasional bit errors) the original transmitted data. In repeater systems the demultiplex operation would be omitted and the detected composite signal could serve as the input for the next troposcatter link.

Because the received signal strength depends on a scattering process that is dependent on fluctuations in the refractive index within the common volume, variations in meteorologic conditions such as temperature and humidity in the common volume produce long-term signal fading. This long-term fading is usually characterized by an hourly median value of received signal power or transmission loss in decibels. Its variation has been found to closely approximate a Gaussian distribution so it is completely described by the hourly median and the standard deviation.

Of course, the scattering process itself results in a fading signal as different scattering paths will add or subtract at the receiver in a random fashion. The individual scattering "blobs" in the common volume tend to be statistically independent so the received signal can be represented in complex notation as the sum of a large number of independent complex random variables. Using the central-limit theorem, the complex received signal is then close to a complex Gaussian process with an envelope, i.e., its magnitude, that follows the Rayleigh probability distribution. If x is the received signal power value and x_m is its median value, the probability that the random variable *x* is less then a critical value x_c is

$$\rho(x < x_c | x_m) = 1 - e^{-x_c \ln 2/x_m}$$

which can be approximated for small values as

$$\rho(x < x_c | x_m) \doteq \frac{x_c \ln 2}{x_m} \quad x_c \ll x_m$$

The fading probability distribution above shows that deep fades have a slope of 10 dB per probability decade. This implies that to achieve short-term availabilities of 99.9% there must be a fade margin protection of about 30 dB. One way to reduce this severe power penalty due to fading is to provide redundant signal paths called *diversity paths*. Different diversity configurations are considered in the next section. In general for *M*th-order diversity (i.e., *M* independent transmission paths between transmitter and receiver), the fading slope reduces to 10/*M* dB per probability decade. Thus a quadruple diversity system requires about $\frac{30}{4} = 7.5$ dB of fade margin protection relative to the 30 dB for the no diversity system at a 99.9% availability.

The advantages of troposcatter systems include

- Realization of long paths beyond the LoS zone
- Use over difficult terrain where installing cable systems is unattractive
- Physical security because prevention of sabotage or catastrophic failure is easy to achieve with a small number of stations
- High immunity to interception or to jamming because of the use of narrow beam antennas
- Provision of a communication service that does not depend on satellite services
- Rapid link installation (about one day)

The disadvantage of tropospheric scatter networks include poorer quality signals and lower data rate capacities than alternative satellite and cable systems.

3. DIVERSITY CONFIGURATIONS

In a multipath fading environment it is common to add or utilize redundant transmission paths so that the fading of one or more paths will not necessarily prevent correct reception of the transmitted information. This redundancy, called *diversity*, is a critical element in the successful operation of troposcatter links. For example, in the Rayleigh fading environment for tropospheric circuits, if an outage level of 1% was acceptable, a no diversity system would require about 11.5 dB more transmit power than a dual-diversity system with two separate receive antennas. Since the antennas and power amplifiers in most troposcatter circuits are near technological limits, diversity is an extremely attractive performance improvement approach. Diversity techniques include frequency using multiple radiofrequency carriers, space using multiple transmit or receive antennas, polarization using orthogonal polarization transmissions,



Figure 1. Dual-diversity configurations: (**a**) dual-frequency (2F) diversity; (**b**) dual space (2S) diversity. (D = duplexer).

and angle using an antenna feedhorn that produces multiple beams.

3.1. Frequency and Space Diversity

The most common diversity configurations use combinations of frequency and space as shown in Figs. 1 and 2. A diamond and bar are used in these figures to denote horizontal and vertical polarization, respectively. Troposcatter circuits are always duplex, so frequency diversity requires a duplexer that allows a transmitter and receiver operating in two different frequency bands to be connected to the same antenna. Transmit powers are on the order of 1 kW and receivers may be required to detect signals at levels near — 100 dBm, a dynamic range of 160 dB. In the dual-space diversity, only one transmit antenna needs to be used but equipment redundancy requires a second power amplifier. For equal power amplifier outputs the dual-space (2S) diversity of Fig. 1b is then 3 dB better in total received power than the dual frequency (2F) diversity of Fig. 1a because the total transmit power is the same but the 2S system receives on two antennas.

A 2S/2F quadruple diversity can be achieved with a combination of frequency and space as shown in Fig. 2a. Under certain conditions [6] a savings of two frequency channels per path can be achieved by replacing the frequency channels with orthogonal polarizations. This configuration can be denoted as quadruple space (4S) because there are four separate space diversity paths. The decorrelation of these paths is due to their spatial separation through the scattering common volume. This configuration is also called dual space/dual polarization (2S/2P) although the polarization serves only to separate the signals to each antenna for receiver processing. Generally polarization diversity with



Figure 2. (a) Quadruple space/frequency (2S/2F) diversity and (b) quadruple space (4S) diversity (D = duplexer).

the same spatial path, e.g., a 2P system with one antenna at each terminal, does not realize sufficient decorrelation to provide a diversity effect.

The antenna spacing in space diversity systems is more conveniently accomplished in the horizontal direction and a separation of 100 wavelengths provides good decorrelation [6] and is widely adopted.

3.2. Angle Diversity

Redundant and decorrelated paths can be realized with multiple beams produced at either the transmitter or receiver antenna. Beams can be separated either horizontally or vertically. The antenna beams can be produced by replacing the feedhorn at the focal point of a parabolic reflector with a multiple feedhorn structure. This method is normally accomplished at the receiver because transmit diversity requires either extra power amplifiers or a division of the available transmitter power.

The selection of a horizontal or vertical displacement of beams depends on the beam correlation for these axes. Beam correlation is inversely related to the common volume power density width on that axis. Horizontally the common volume "size" is limited to approximately the beamwidth. In the vertical dimension the common volume "size" extends beyond the beamwidth limit with a scattering angle dependence of $\theta^{-11/3}$, where θ is generally larger than the beamwidth. Thus the beam correlation for conventional narrowbeam antennas is smaller in



Figure 3. Dual-space/dual-angle (2S/2A) diversity.

the vertical direction. An experimental angle diversity system [7] at 5 GHz used a two-beam vertical splay with a feedhorn design that produced a beam separation or "squint angle" of approximately 1.3 beamwidths compared to a theoretical optimum of about 1 beamwidth. The loss in performance in this experimental study with the slightly larger squint angle was determined to be between 0.1 and 0.4 dB. The configuration of the experimental dualspace/dual-angle (2S/2A) system is shown in Fig. 3.

A major advantage of angle diversity is that it can be used to replace frequency diversity in a 2S/2F system to produce a 2S/2A system that saves two frequency channels per path and results in better performance. In a vertical dual-angle system, there is a performance loss of approximately one-half the decibel value of the loss associated with the larger scattering angle of the elevated beam and a small loss because of beam correlation. These combined losses in angle diversity are generally less than the 3 dB received power advantage relative to frequency diversity. Troposcatter systems require two power amplifiers for equipment redundancy, but in angle diversity they operate in the same frequency band (see Fig. 3), producing a potential received power 3 dB larger than frequency diversity at each diversity receiver. In addition to this short-term advantage, the 2S/2A system is superior to the 2S/2F system because long-term variations are decorrelated in the two antenna beams; thus troposcatter systems should have adaptively compensated takeoff angles. The experimental results reported [7,23] confirm both the short- and long-term advantages of the angle diversity system.

4. TRANSMISSION PARAMETER PREDICTION

Troposcatter propagation is possible in the frequency range from a lower limit determined by antenna size of a few hundred megahertz (MHz) up to about 10,000 MHz where water vapor losses become excessive. Practical link distances range from 50 to 500 km and channel capacities are typically 60–120 telephone channels in analog systems and up to about 12 Mbps in digital systems. These systems must cope with long-term (hourly) and short-term (seconds) variations in received signal strength and multiple paths between the transmitter and receiver that produce self-interference. Predictions of transmission loss and multipath characteristics are essential in the communication link design.

4.1. Transmission Loss

Comprehensive methods for predicting cumulative probability distributions of transmission path loss as a function of radiofrequency and distance over any type of terrain and in different climate regions were produced by the Comite Consultatif International Radio (CCIR) [9] and in the United States by the National Bureau of Standards (NBS) [10]. The detailed point-to-point predictions depend on propagation path geometry, atmospheric refractivity near the center of the earth, and specified characteristics of antenna directivity.

Path loss can be broken into two major components: (1) a basic transmission loss for a system with hypothetical loss-free isotropic antennas and (2) an antenna gain loss that accounts for the reduction in the illuminated portion of the scattering common volume when the beamwidth is reduced. The basic transmission loss depends on the radio frequency, the distance, and the scattering angle between the centerlines of the transit and receive beams at their intersection in the scattering region. The scattering angle, θ , depends on the pathlength d and the transmit and receive elevation angles, θ_{et} and θ_{er} , respectively. For $\theta d < 2$, the scattering angle can be approximated by

$$\theta = \frac{d}{a} + \theta_{et} + \theta_{er}$$

where a = kR is the effective earth radius to account for bending due to a decrease in refractive index with height. Typical troposcatter calculations multiply the earth radius *R* by a factor of *k* that is $\frac{4}{3}$ [5].

Except for the small correction factors, the NBS method [10] computes the basic transmission loss in dB as

$$L_{bsr} = 30\log f - 20\log d + F(\theta d)$$

where the frequency is in MHz, the path distance is in kilometers, and F is an empirical attenuation function that depends on surface refractivity N_s . The surface refractivity at a height h km above sea level is related to the refractive index n_0 at sea level by

$$N_s = (n_0 - 1) \times 10^6 \times e^{-0.1057h}$$

The attenuation function is provided graphically in Fig. 9.1 in Ref. 10 for various values of N_s . At a nominal value $N_s = 301$, this function can be approximated for $0.017\theta d \le 10$ by

$$F(\theta d) = 30\log(\theta d) + 0.332\theta d + 135.82$$

The loss associated with the scattering process can be appreciated by a comparison with free space loss:

$$L_{\rm FS} = 32.446 + 20\log f + 20\log d$$

Thus at 1000 MHz and 200 km the free-space loss is 138.5 dB, whereas the corresponding troposcatter loss with a nominal scattering angle of 0.01 rad is 188.8 dB. This additional 50 dB loss has to be overcome with large antennas and power amplifiers.

The second component of troposcatter transmission loss, antenna gain loss, occurs with high-gain antennas. An increase in gain increases the illumination in the scattering common volume proportionally to the square of the beamwidth, but the number of illuminated scatterers in the common volume of beam intersection decreases as the cube of the beamwidth so there is an overall antenna gain loss proportional to the beamwidth. For antenna gains less than 55 dB and for gains that are not very different for the two antennas, the antenna gain loss in dB can be expressed as [5]

$$\Delta g = 0.07 \exp\{0.055(G_t + G_r)\}$$

where G_t and G_r are the free-space antenna gains in decibels (dB) of the transmitter and receiver, respectively. A more exact calculation of gain loss can be obtained by numerical integration of the scattering common volume defined by the antenna pattern distribution. Parl [11] has developed closed-form results for the path loss and its component antenna gain loss that agree well with the numerical integration results. Antenna gain loss is also called *aperture-to-medium coupling loss* and was the subject of numerous analyses [e.g., 12,13].

4.2. Variability

The transmission loss calculation for a troposcatter communication link provides a prediction of the median value over a measurement period that might be months or a year. Within that measurement period the received signal will undergo variations due to changes in the troposphere meteorologic conditions. Slow changes in temperature and humidity result in corresponding variations in refractive index that can be represented by hourly median values during the measurement period. The distribution of these hourly medians expressed in a dB measure has been found to be approximately Gaussian. The prediction of transmission loss corresponds to the median of this distribution and the standard deviation in dB completes the statistical definition of this long-term variation.

In the NBS method the prediction of the longterm variation is accomplished from a set of empirical curves [14] of variability $V(\rho, \theta)$ in dB as a function of the scatter angle θ and the link availability that the transmission median loss is not exceeded $\rho\%$ of the time. For example, at $\rho = 99.9\%$ the standard deviation of the Gaussian variability distribution is V/3.1.

It was recognized by system designers that the prediction of the transmission loss is subject to uncertainty. This uncertainty was expressed as a variance $\sigma_c^2(\rho)$ dependent on the link availability ρ . The system would be designed with a service probability $F(\tau)$, where τ is the standard normal deviate and the predicted transmission loss would be increased by $\tau \sigma_c(\rho)$. Empirical

curves for the prediction uncertainty $\sigma_c(\rho)$ are given in Ref. 14.

There is also a variation within the hour due to changes in the pathlengths associated with individual scatterers in the common volume of the intersection of the antenna beams. This short term fading of the received signal envelope is due to multipath and has been found to follow a Rayleigh distribution. The system designer will normally require an hourly median value of transmission loss that includes a fade margin to ensure satisfactory performance under short term Rayleigh fading conditions. The Rayleigh distribution was defined and fade margins are discussed in Section 2.

4.3. Multipath Dispersion

The delay difference between the shortest and longest paths from the transmitter to receiver through the scattering common volume is the delay dispersion. In an analog system, telephone channels are at different frequencies in the composite transmitted signal. A delay difference might produce additive interference for one channel and subtractive interference for another channel. This type of interference and the nonlinear demodulation of frequency modulation produces intermodulation distortion in analog systems.

In a digital system, delay dispersion will cause interference between adjacent symbols, namely, intersymbol interference (ISI). If this ISI is not taken into account by the digital demodulation process, serious performance degradation will result. Adaptive receivers exploit this multipath phenomenon as a form of signal transmission redundancy; that is, the same transmitted information is associated with multiple delay paths.

The basic prediction method for delay dispersion is to calculate the path distances and their corresponding delay through the common volume. Sunde [15] uses an approximate formula to obtain the maximum departure from the mean transmission delay. Sunde also used this result to predict intermodulation distortion in analog systems [16].

In a channel model developed by Bello [17], a onedimensional approximation to the common volume was used to derive estimates of the root-mean-square (RMS) value of multiple delay spread. Multipath measurements by Sherwood and Suyemoto [18] found two-sided RMS multipath spread, 2σ , to be considerably wider than predicted by the Bello model. They also observed considerable variation in measured 2σ values but no significant correlation with variables such as path loss, surface refractivity, and effective earth radius. Collin [19] measured the cumulative distribution of the 2σ multipath spread and approximated the distribution by a lognormal law. From these empirical data, 99% 2σ values were predicted and compared favorably with data on 0.9- and 4.8-GHz links.

Multipath spread variation has two major causes: changes in the effective earth radius and layering due to turbulence that modify the refractive index height profile within the scattering common volume. In the absence of empirical data on this height profile, a worst-case

$$\tau = d(\alpha_{t1}\alpha_{t0} - \alpha_{r1}\alpha_{r0})$$

maximum delay spread τ can be calculated [20] as

where d is the pathlength, c is the speed of light, and α_{t1} , $\alpha_{t0}(\alpha_{r1}, \alpha_{r0})$ are the maximum and minimum takeoff angles at the transmitter (receiver) measured from the straight line bisecting the transmitter and receiver to the lowest and highest points in the scattering volume, respectively. The worst-case multipath spread can be used to ensure that intermodulation noise does not limit capacity in an analog system and that adaptive systems are sufficiently robust in digital systems. For predicting nominal performance the yearly median 2σ value can be calculated [20] by three-dimensional integration for a fixed refractive index within the scattering common volume.

5. ANALOG SYSTEM PERFORMANCE

In an analog troposcatter link typically 60-120 4-kHz telephone channels are frequency division multiplexed to provide a baseband signal that is subsequently frequency modulated. Link performance can be determined by computing the median value of the voice channel signal-tonoise ratio (SNR) or an outage probability representing the fraction of time the SNR is below a critical threshold. The latter criterion is more meaningful in typical applications where the telephone call includes multiple troposcatter links. For small outage probabilities, the network outage probability is simply the sum of the link probabilities.

An analysis adapted from Ref. 21 is presented here for analog systems that include the effects of both signal-level variations and multipath delay distortion. The measure of performance selected is the signal-to-noise ratio r in a voice channel. The noise is defined to include the effects of thermal noise due to signal level variations and intermodulation (IM) noise due to multipath delay variations. The short-term probability is

$$P(\overline{C}, S) = \operatorname{prob}(r < r_c) \tag{1}$$

where r_c is a critical value of r, the voice channel SNR, \overline{C} is the average received unmodulated carrier power, and S is the 2σ multipath delay spread.

The effects of thermal noise and path intermodulation disturbance add on a noise power basis, so that we have

$$r = (x^{-1} + y^{-1})^{-1}$$
(2)

where x is the signal-to-thermal noise ratio and y is the signal-to-path IM ratio.

5.1. Thermal Noise Effects

The signal-to-thermal noise ratio has a mean value \overline{x} from FM theory [22]:

$$\overline{x} = \frac{\overline{C}}{N_0 B} \left(\frac{f_d}{f_1}\right)^2 \frac{B}{N} g_p, \frac{\overline{C}}{N_0 B} > T_{\rm FM}$$
(3)

where *B* is the IF bandwidth, f_d is the RMS deviation of the voice channel signal, f_1 is the frequency location of

the voice channel in the multiplex format, b is the voice channel bandwidth, and g_p is the preemphasis factor for the voice channel location. Equation (3) is valid provided the carrier-to-noise ratio $\text{CNR} = \overline{C}/N_0 B$ is greater than the FM threshold, T_{FM} . With threshold extension T_{FM} occurs at a CNR value of about 7 dB. Below FM threshold the signal-to-noise ratio drops rapidly with CNR. The random variable x can then be approximated by a piece-wise linear function of CNR in dB. For a general demodulator, we define

$$x = g(C) = \begin{cases} g_1(C) & \frac{C}{N_0 B} \ge T_{\rm FM} \\ g_2(C) & \frac{C}{N_0 B} \prec T_{\rm FM} \end{cases}$$
(4)

where C is the instantaneous carrier power and \overline{C} is its short-term mean. The linear functions g_1 is given by Eq. (3) and g_2 is obtained from FM demodulator characteristics below threshold.

Since C is a short-term random variable, so is the signal-to-thermal noise ratio x. With optimum combining, C has a gamma PDF of order D, where D is the number of diversities. The probability distribution for C, that is, the probability that the carrier power after diversity combining is less than C, is given by

$$F_C(C) = e^{-DC/\overline{C}} \quad \sum_{I=D}^{\infty} \frac{(DC/\overline{C})^I}{I!}$$
(5)

When path IM is negligible, the outage probability can be computed from knowledge of $F_C(A)$ by

$$P(\overline{C}, O) = F_C(g^{-1}(r_c)) \tag{6}$$

where $g^{-1}(x)$ is the inverse relation defined by (4).

5.2. Path Intermodulation Effects

In general the analysis becomes much more difficult when multipath delay effects have to be considered. Multipath delay variations result in an intermodulation noise when the distorted signal passes through the FM discriminator of an analog system. The performance due to this effect is summarized in a signal-to-intermodulation noise ratio (SINR). One commonly used approach for calculating the median SINR ratio is due to Sunde [16]. His results are expressed in terms of the noise power ratio (NPR) [2]. When this ratio is converted to voice channel signal-tointerference ratio, the result for the SINR ratio is

$$y(\Delta) = \left(\frac{f_d}{B}\right)^2 \frac{f_1}{b} \frac{1}{GH(\gamma)} \tag{7}$$

where *B* is the baseband bandwidth, *G* is the preemphasis factor, and $H(\gamma)$ is the degradation due to phase distortion γ . The preemphasis factor used by Sunde was G = 0.192. The phase distortion γ is a function of the baseband RMS deviation F_r and a delay departure Δ from the mean transmission delay:

$$\gamma = 8F_r^2 \Delta^2 \tag{8}$$

2698 TROPOSPHERIC SCATTER COMMUNICATION

In Sunde's original work the function $H(\gamma)$ is proportional to γ^2 (i.e., Δ^4) for small phase distortion values. The saturation effect is due to the use of only the quadratic phase distortion term and omission of higher-order terms. Since for large phase distortion, the multipath effect in an FM system should be proportional to the multipath power, it is more reasonable to have $H(\gamma)$ asymptotically be proportional to γ (i.e., Δ^2). This leads to a modified version for the phase distortion function of the form

$$H(\gamma) = \begin{cases} \gamma^2 & \gamma < \frac{1}{2} \\ \frac{1}{2}\gamma & \gamma > \frac{1}{2} \end{cases}$$
(9)

This function is identical to Sunde's below $\gamma = \frac{1}{2}$ and is approximately tangent to the saturation portion around $\gamma = 1$. Equations (7) and (9) thus allow one to compute the SINR value in terms of a multipath delay Δ and the FM system parameters. Unfortunately, little is known about the short-term distribution of Δ for which *S* is a statistic.

In Ref. 21 a short-term model is proposed that treats Δ as a random variable and the probability distribution F_{Δ} is derived as an implicit function of the multipath spread S.

The short-term outage probability due to path intermodulation effects alone can then be found from this multipath delay distribution and the relation between signal-to-interference ratio y(d) as a function of delay by Eqs. (7) and (8). The outage probability for path IM alone is

$$P(\infty, S) = F_{\Delta}(\Delta(r_c)) \tag{10}$$

where $\Delta(y)$ is the inverse relation to Eq. (7) for delay in terms of SINR ratio.

5.3. Combined Thermal Noise and Path IM Effects

Given the probability distributions for the signal-tothermal noise ratio x and the signal-to-intermodulation noise ratio y, it is a straightforward but tedious task to compute the probability distribution of the total SNR [Eq. (2)] for independent x and y values. The result is

$$p(C, S) = \operatorname{prob}(r < r_c)$$

= $\int_0^{r_c} dF_c(g^{-1}(x)) \int_0^{(r_c^{-1} - x^{-1})^{-1}} dF_\Delta(\Delta(y))$ (11)

In most cases however the outage probability is dominated by one effect or the other. Thus, a reasonable approximation is to use

$$p(\overline{C}, S) \doteq p(\overline{C}, O) + p(\infty, S)$$
(12)

as the short-term outage probability.

Empirical evidence [10,19] suggests that the long-term fluctuations in \overline{C} and S can be described by a lognormal density function. Also the joint process appears to be sufficiently decorrelated that performance estimates can assume independence in terms of the means and standard deviation of \overline{C} and S. Estimates of these parameters for \overline{C} can be obtained in Ref. 10 and for S in Refs. 17–19.

6. DIGITAL SYSTEMS

The multipath delay spread limits the channel capacity that can be achieved in analog systems. Only transmission bandwidths less than the reciprocal of this multipath delay spread can be achieved. Signals of larger bandwidths become distorted due to the multipath dispersion. In FM systems this dispersion causes intermodulation noise after detection.

With digital signal formats, adaptive methods can be used to measure the multipath structure and exploit it as an extra form of diversity to improve performance. Unlike the capacity of analog systems, the capacity of digital systems is not restricted by the multipath delay spread. From a network viewpoint, fades in tandem digital links do not have a cumulative effect because the signal can be regenerated at each node.

Adaptive troposcatter systems have been demonstrated that are efficiently able to detect digital signals perturbed by a fading channel medium while tracking the fading variations. Applicability of adaptive signal processing techniques is critically dependent on whether the rate of fading is slower than the rate of signaling. As discussed below, troposcatter radio links can be considered to be slow-fading multipath channels.

6.1. Slow-Fading Multipath Channels

For digital communication over troposcatter radio links, an attempt is made to maintain transmission linearity; thus the receiver output should be a linear superposition of the transmitter input plus channel noise. This is accomplished by operation of the power amplifier in a linear region or, with saturating power amplifiers, by using constantenvelope modulation techniques. For linear systems, multipath fading can be characterized by a transfer function of the channel H(f;t) that is the frequency domain response at a carrier of 0 Hz as a function of time t. Let t_d and f_d be the decorrelation separations in the time and frequency variables, respectively. If t_d is a measure of the time decorrelation in seconds, then

$$\sigma_t = rac{1}{2\pi t_d} \, \mathrm{Hz}$$

is a measure of the fading rate or bandwidth of the random channel. The quantity σ_t is often referred to as the *Doppler spread* because it is a measure of the width of the received spectrum when a single sine wave is transmitted through the channel. The dual relationship for the frequency decorrelation f_d in hertz suggests that a delay variable

$$\sigma_{f} = rac{1}{2\pi f_{d}} \; \mathrm{s}$$

(in seconds) defines the extent of the multipath delay. The quantity σ_f is approximately equal to the RMS multipath delay spread Φ that represents the RMS width of the received process in the time domain when a single impulse function is transmitted through the channel.

Typical values of Doppler and delay spread for troposcatter communication are around 1 Hz and 100 ns, respectively.

The spreads can be defined as moments of spectra in a channel model [24] that assumes wide-sense stationary (WSS) in the time variable and uncorrelated scattering (US) in the multipath delay variable. This WSSUS model and the assumption of Gaussian statistics for H(f;t)provide a statistical description in terms of a single twodimensional correlation function of the random process H(f;t).

This characterization has been quite useful and accurate for a variety of radio link applications. However the stationary and Gaussian assumptions are not necessary for the utilization of adaptive signal processing techniques on these channels. What is necessary is first that sufficient time exists to "learn" the channel characteristics before they change, and second, that decorrelated portions of the frequency band be excited such that a diversity effect can be realized. These conditions are reflected in the following two relationships in terms of the previously defined channel factors, the data rate R and the bandwidth B.

$R \text{ (bps)} \gg \sigma_t(\text{Hz})$ learning requirement

B (Hz) $\geq f_d$ (Hz) diversity requirement

The learning requirement insures that the channel remains approximately fixed for an interval containing many received bits. The energy in these received bits provides a basis for measurement of the channel characteristics. If $R \sim \sigma_t$, the channel would change before significant energy for measurement purposes could be collected. The signal processing techniques in an adaptive receiver do not necessarily need to measure the channel directly in the optimization of the receiver, but the requirements on learning are approximately the same. If only information symbols are used in the sounding signal, the learning mode is referred to as decision-directed. When digital symbols known to both the transmitter and receiver are employed, the learning mode is called referencedirected. Adaptation of troposcatter receivers with no wasted power for sounding signals can be accomplished using the decision-directed mode. This is possible because of the small number of adaptation parameters, the continuous nature of the communication, and the high likelihood that receiver decisions are correct.

As described in Section 2, diversity in fading applications is used to provide redundant communications channels so that when some of the channels fade, communication will still be possible over the others that are not in fade. These diversity techniques are sometimes called explicit diversity because of their externally visible nature. An alternate form of diversity is termed implicit diversity because the channel itself provides redundancy. In order to capitalize on this implicit diversity for added protection, receiver techniques have to be employed to correctly assess and combine the redundant information. The potential for implicit frequency diversity arises because different parts of the frequency band fade independently. Thus, while one section of the band may be in a deep fade, the remainder can be used for reliable communication. However, if the transmitted bandwidth Bis small compared to the frequency decorrelation interval

 f_d , the entire band will fade and no implicit diversity can result. Thus, the second requirement $B \ge f_d$ must be met if an implicit diversity gain is to be realized. In diversity systems a little decorrelation between alternative signal paths can provide significant diversity gain. Thus it is not necessary for $B \gg f_d$ in order to realize implicit frequency diversity gain, although the implicit diversity gain clearly increases with the ratio B/f_d . Note that the condition $R \ll B \ge f_d$ does not preclude the use of implicit diversity because a bandwidth expansion technique can be used in the modulation process to spread the transmitted information over the available bandwidth B. Most digital troposcatter applications, however, are high-rate where Rand B are about the same.

The implicit diversity effect described here results from decorrelation in the frequency domain in a slow-fading $(R \gg \sigma_t)$ application. This implicit frequency diversity can in some circumstances be supplemented by an implicit time diversity effect, which results from decorrelation in the time domain. In fast-fading applications $(R > \sigma_t)$ redundant symbols in an error-correcting coding scheme can be used to provide time diversity provided the codeword spans more than one fade epoch. In our slowfading application this condition of spanning the fade epoch can be realized by interleaving the codewords to provide large time gaps between successive symbols in a particular codeword. The interleaving process requires the introduction of signal delay longer than the time decorrelation separation t_d . Methods of realizing implicit diversity gain with decoding delays as short as $\frac{1}{4}$ second have been studied [25]. However in many practical applications that require transmission of digitized speech over multiple troposcatter links, the required time delay is unsatisfactorily long for two-way speech communication. For this reason implicit frequency rather than time diversity techniques are used in existing systems. The receiver structures to be discussed next are applicable to situations where implicit frequency diversity can be realized.

6.2. Digital Receivers

When the transmitted symbol rate is on the order of the frequency decorrelation interval of the channel, the frequencies in the transmitted pulse will undergo different gain and phase variations resulting in reception of a distorted pulse.

Although there may have been no intersymbol interference (ISI) at the transmitter, the pulse distortion from the channel medium will cause interference between adjacent samples of the received signal. In the time domain, ISI can be viewed as a smearing of the transmitted pulse by the multipath causing overlap between successive pulses. The condition for ISI can be expressed in the frequency domain as

$$T^{-1}(\mathrm{Hz}) \ge f_d(\mathrm{Hz})$$

or in terms of RMS multipath delay spread as

$$T(\text{seconds}) \leq 2\pi\sigma(\text{seconds})$$

2700 TROPOSPHERIC SCATTER COMMUNICATION

Bandwidth limitations and SNR efficiency in the presence of multipath fading make quadrature phase shift keying (QPSK) a common choice for modulation. Since the bandwidth of a QPSK signal is at least on the order of the symbol rate T^{-1} Hz, there is no need for bandwidth expansion under ISI conditions in order to provide signal occupancy of decorrelated portions of the frequency band for implicit diversity. However, it is not obvious whether the presence of the intersymbol interference can wipe out the available implicit diversity gain. It has been established that adaptive receivers can be used that cope with the intersymbol interference and in most cases wind up with a net implicit diversity gain. These receiver structures fall into three general classes: matched filters, equalizers, and maximum likelihood detectors.

6.2.1. Matched Filters. A matched filter has characteristics that "match" the characteristics of the combined transmitter and channel. If the combined frequency response is denoted H(f), the matched filter is $H^*(f)$. In the time domain the combined impulse response can be denoted as h(t). The matched filter impulse response is an inverted and conjugated response: $h^*(-t)$. The matched filter can also be realized by correlating the received signal with the combined impulse response. Since the channel part of the combined response is changing and unknown to the receiver, adaptation is required. Also the process of matched filtering increases the total impulse response thus increasing intersymbol interference (ISI) effects. A reduction in the ISI can be realized by transmitting a shorter pulse and leaving a time gap before the next pulse begins.

Figure 4 illustrates an adaptive matched filter based on these principles. Data modulation is stripped from the received signal by multiplying by previous decisions. The resulting received pulse is then averaged in a recirculating delay line to produce an estimate of h(t). This channel estimate is correlated with the received signal to complete the matched filter operation. The



Figure 4. Adaptive matched filter for QPSK system.

adaptive matched filter shown in Fig. 4 will degrade when there is intersymbol interference between received pulses because the averaging process would add overlapped pulses incoherently. When the multipath spread is less than the symbol interval, this condition can be alleviated by transmitting a time-gated pulse whose OFF time is approximately equal to the width of the channel multipath. The multipath causes the gated transmitted pulse to be smeared out over the entire symbol duration but with little or no intersymbol interference. Because the multipath components are adaptively combined an implicit diversity [26] effect is obtained. In a configuration with both explicit and implicit diversity, moderate intersymbol interference can be tolerated because the diversity combining adds signal components coherently and ISI components incoherently.

Because the OFF time of the pulse cannot exceed 100%, this approach is clearly data-rate-limited for fixed multipath conditions. In addition, the time gating at the transmitter results in an increased bandwidth that may be undesirable in a bandwidth-limited application. The power loss in peak power limited transmitters due to time gating can be partially offset by using two carrier frequencies with independent data modulation [27]. This technique was successfully developed by Raytheon for use in the U.S. Air Force tactical troposcatter system, AN/TRC-170. In this system a V2 model provides a 2S/2F diversity configuration with traffic capability of up to 120 voice channels and a range between 60 and 140 mi. A smaller V3 model uses a dual-frequency configuration. The AN/TRC-170 has hundreds of units in the U.S. military inventory, it has been sold to other countries, and it continues to be used in tactical military applications.

6.2.2. Adaptive Equalizers. Adaptive equalizers use linear filter subsystems with electronically adjustable parameters that are controlled in an attempt to combine multipath components and compensate for intersymbol interference. Tapped delay-line filters are a common choice for the equalizer structure as the tap weights provide a convenient adjustable parameter set. Adaptive equalizers have been widely employed in telephone channel applications [28] to reduce ISI effects due to channel filtering. In a fading multipath channel application, the equalizer can provide three functions simultaneously: noise filtering, matched filtering for explicit and implicit diversity, and removal of ISI. These functions are accomplished by adapting a tapped delay-line filter (TDF) to force an error measure to a minimum. By designing the error measure to include the degradation due to correlated noise, ISI, filtering, and improper diversity combining, the TDF will minimize their combined effects.

A linear equalizer (LE) is defined as an equalizer that linearly filters each of the N explicit diversity inputs. An improvement to the LE is realized when an additional filtering is performed on the detected data decisions. Because it uses decisions in a feedback scheme, this equalizer is known as a *decision-feedback equalizer* (DFE).

The operation of a matched filter receiver, an LE, and a DFE can be compared from examination of the received



Figure 5. Received pulse sequence.

pulsetrain example of Fig. 5. The binary modulated pulses have been smeared by the channel medium producing pulse distortion and interference from adjacent pulses. Conventional detection without multipath protection would integrate the process over a symbol period and decide whether a +1 was transmitted if the integrated voltage is positive and -1 if the voltage is negative. The pulse distortion reduces the margin against noise in that integration process. A matched filter correlates the received waveform with the received pulse replica, thus increasing the noise margin. The intersymbol interference arises from both future and past pulses in these radio systems since the multipath contributors near the mean path delay normally have the greatest strength. This ISI can be compensated for in a linear equalizer by using properly weighted time-shifted versions of the received signal to cancel future and past interferers. The DFE uses time-shifted versions of the received signal only to reduce the future ISI. The past ISI is canceled by filtering past detected symbols to produce the correct ISI voltage from these interferers. The matched filtering property in both the LE and DFE is realized by spacing the taps on the received signal TDFs at intervals smaller than the symbol period.

The DFE is shown in Fig. 6 for the Nth-order explicit diversity system. A forward filter (FF) TDF is used

for each diversity branch to reduce correlated noise effects, provide matched filtering and proper weighting for explicit diversity combining, and reduce ISI effects. After diversity combining, demodulation, and detection, the data decisions are filtered by a backward filter TDF to eliminate intersymbol interference from previous pulses. Because the backward filter compensates for this "past" ISI, the forward filter need only compensate for "future" ISI.

A decision-directed error signal for adaptation of the DFE is shown as the difference between the detector input and output. Qualitatively one can see that if the DFE is well adapted, this error signal should be small. Referencedirected adaptation can be accomplished by multiplexing a known bit pattern into the message stream for periodic adaptation.

When error propagation due to detector errors is ignored, the DFE has the same or smaller meansquare error than the LE for all channels [29]. The error propagation mechanism has been examined by a Markov chain analysis [30] and shown to be negligible in practical fading applications. Also in an Nth-order diversity application, the total number of TDF taps is generally less for the DFE than for the LE. This follows because the former uses only one backward filter after combining of the diversity channels in the forward filter.

The performance of a DFE on a fading channel can be predicted [31] using a transformation technique that converts implicit diversity into explicit diversity and treats the ISI effects as a Gaussian interferer. An example of this calculation for a quadruple diversity system with $2\sigma/T = 0$ and 0.5 is given in Ref. 32. The average probability of error is poorest when there is no multipath spread because there is no implicit diversity. The performance difference between a three-tap (3T) forward filter and an ideal matched filter was also shown to be less than a dB. Thus with moderate $2\sigma/T$ values the DFE performance is close to optimum.



Figure 6. Decision-feedback equalizer, Nth-order diversity.



Figure 7. Diversity DFE performance versus. multipath spread QPSK modulation at data rates of 3.2, 6.4, and 9.7 Mbps: three forward filter taps at T/2 spacing, three backward filter taps, Eb/N0 in dB/diversity.

A DFE modem was developed [33] with data rates up to 12.5 Mbps for application on troposcatter channels with up to four orders of diversity. This DFE modem uses only a three-tap forward filter TDF and a three-tap backward filter TDF. Extensive simulator and field tests [31,33,34] have shown that implicit diversity gain is realized over a wide range of actual conditions while ISI effects are mostly eliminated. Figure 7, reproduced from Fig. 6.5 of Ref. 34, summarizes simulated performance as a function of 2σ multipath spread and signal-to-noise ratio. The improvement in bit error rate when $2\sigma/T$ increases from zero is due to implicit diversity. For larger values of $2\sigma/T$ approaching 2, the ISI begins to dominate the implicit diversity effect. Measured results agree well with the predicted performance from Ref. 31.

An 8-Mbps version of the DFE modem is produced commercially by Comtech Systems Inc., Orlando FL and has been incorporated in both military and civilian tropospheric scatter radio systems. Applications of the latter include ocean oil platforms and a backbone angle diversity system between islands in the Bahamas for cellular telephone service.

6.2.3. Maximum-Likelihood Detectors. The DFE is not optimum for all channels with respect to bit error probability. By considering intersymbol interference as a conventional code defined on the real line (or complex line for bandpass channels), maximum-likelihood sequence estimation algorithms have been derived [35,36] for the linear modulation channel. These algorithms provide a decoding procedure for receiver decisions that minimize the probability of sequence error. A maximum-likelihood

sequence estimator (MLSE) receiver requires matched filters for each diversity channel and a combiner. After these filtering and combining operations, a trellis decoding technique is used to find the most likely transmitted sequence.

The MLSE algorithm works by assigning a state for each intersymbol interference combination. Because of the one-to-one correspondence between the states and the ISI, the maximum-likelihood source sequence can be found by determining the trajectory of states.

If some immediate state is known to be on the optimum path, then the maximum-likelihood path originating from that state and ending in the final state will be identical to the optimal path. If at time n, each of the states has associated with it a maximum likelihood path ending in that state, it follows that sufficiently far in the past history will not depend on the specific final state to which it belongs. The common path history is the maximumlikelihood state trajectory [36].

Since the number of ISI combinations and thus the number of states is an exponential function of the multipath spread, the MLSE algorithm has complexity that grows exponentially with multipath spread. The equalizer structure exhibits a linear growth with multipath spread. In return for additional complexity, the MLSE receiver results in a smaller (sometimes zero) intersymbol interference penalty for channels with isolated and deep frequency selective fades. However in many applications where high orders of diversity are employed, these deep selective frequency fades do not occur frequently enough to significantly affect the average error probability [33]. Partly for these reasons the MLSE technique for troposcatter applications never advanced beyond the experimental model phase.

7. PRESENT APPLICATIONS OF TROPOSCATTER SYSTEMS

Satellite and cable systems can better provide for large traffic capacities or generally high data rates. However, for links with total data rate of about 10 Mbps, the fixed cost of a digital troposcatter system-satellite transponder lease cost compares favorably enough that troposcatter systems continue to be implemented. Ocean oil platforms and backbone systems for cellular telephone service have been cited here as examples. The utility of troposcatter systems in a tactical military application is also still in evidence. One also notes applications in smaller countries that do not have the technology or finances to implement their own satellite system and are reluctant to lease satellite services that could be monitored or terminated for political reasons. Finally there are applications in disaster relief and in data internet services to remote locations.

BIOGRAPHY

Peter Monsen received a B.S. in electrical engineering from Northeastern University in 1962, a M.S. in operations research from Massachusetts Institute of Technology in 1963, and the Doctor of Engineering Science in electrical engineering from Columbia University in 1970.

From 1964 to 1966 Dr. Monsen served as a Lieutenant in the U.S. Army at the Defense Communications Agency. From 1966 to 1972 he was employed by AT&T Bell Laboratories as a supervisor of a Transmission Studies Group working on fading-channel characterization and adaptive equalization. During this period, the optimum decision-feedback equalizer and a method to adapt it on a time-varying radio channel were developed. From 1972 to 1984 he was employed by SIGNATRON, Inc., working on a wide range of signal processing techniques including adaptive equalization, diversity combining, dual polarization utilization, and interference cancellation. This work led to the conception and development of a 12.6 Mb/s adaptive equalizer troposcatter modem for Defense Communication System and NATO applications with a first operational link between Berlin and West Germany.

In 1984, PM Associates was founded through which consulation for a broad range of telecommunication companies has been provided. Independently, Dr. Monsen has recently submitted three patent applications on multiple access systems with unity reuse factor.

BIBLIOGRAPHY

- 1. Van Der Pool and H. Bremmer, The diffraction of electromagnetic waves from an electrical point source around a finitely conducting sphere, with applications to radiotelegraphy and the theory of the rainbow, *London*, *Edinburg*, *Dublin Philos. Maga. J. Sci.* **24**(164): 141 (July 1937) (series 7).
- 2. P. F. Panter, Communications System Design-Line of Sight and Troposcatter Systems, McGraw-Hill, New York, 1972.

- W. E. Gordon, Radio scattering in the troposphere, *Proc. IRE* 43: 23 (Jan. 1955).
- 4. F. duCartel, Propagation Tropospherique et Faisceaux Hertiziens Transhorizon, Cheron, Paris, 1961.
- G. Roda, *Troposcatter Radio Links*, Artech House, Norwood, MA, 1988.
- R. Larsen, Quadruple space diversity in troposcatter systems, Marconi Rev. 28-55 (first quarter 1980).
- 7. G. Krause and P. Monsen, Results of an angle diversity test experiment, *Natl. Telecommunications Conf. Record*, 1978.
- 8. P. Monsen and S. Parl, Adaptive Antenna Control (AAC) Program, Final Report, ADA091764, Aug. 1980.
- CCIR, The Concept of Transmission Loss in Studies of Radio Systems, documents of the Xth Plenary Assembly, ITU, Geneva, Vol. III, Recommendation 341, 1963.
- P. L. Rice, A. G. Longley, K. A. Norton, and A. P. Barsis, *Transmission Loss Predictions for Tropospheric Communi cation Circuits*, Tech. Note 101, Vols. I–II, National Bureau of Standards, Boulder, CO, 1965.
- S. Parl, New formulas for tropospheric scatter path loss, Radio Sci. 14(1): 49-57 (Jan.-Feb. 1979).
- L. P. Yeh, Experimental aperture-to-medium coupling loss, Proc. IRE 50: 203 (1962).
- L. Boithias and J. Battesti, Etude experimental de la baisse du gain d'antenna dans less liaisons transhorizon, Ann. Telecommun. 19(9-10): 221,229 (1964).
- A. F. Barghausen et al., Ground Telecommunication Performance Standards, NBS Report 6767, Boulder, CO, Part 5 of 6, Tropospheric Systems, Chaps. 5 and 6, 15 June 1961.
- 15. E. D. Sunde, Digital troposcatter transmission and modulation theory, *Bell Syst. Tech. J.* 143-214 (Jan. 1964).
- E. D. Sunde, Intermodulation distortion in analog FM troposcatter systems, *Bell Syst. Tech. J.* 399–435 (Jan. 1964).
- P. A. Bello, A troposcatter channel model, *IEEE Trans.* Commun. Technol. COM-17: 130-137 (April 1969).
- A. Sherwood and L. Suyemoto, *Multipath Measurements over Troposcatter Paths*, Mitre Report MTP-170, Bedford, MA, April 1976.
- C. Collin, Empirical Evaluation of the Correlation Bandwidth in Troposcatter Paths, Revue Technique Thomson-CSF, Vol. 11, No. 3, Sept. 1979.
- P. Monsen et al., Digital Troposcatter Performance Model: Final Report, Signatron, Inc., Lexington, MA, par. 2.5.5, Dec. 1983.
- P. Monsen, Analog and digital transmission performance models for troposcatter links, AGARD Conf. Electromagnetic Wave Propagation Panel Symp., Copenhagen, Denmark, May 1982.
- 22. E. D. Sunde, Communications Systems Engineering Theory, Wiley, New York, 1969, Chap. 4.
- P. Monsen and J. Eschle, Test Report Adaptive Antenna Control Program, ADA04416, U.S. Army, Ft. Monmouth NJ, Sept. 1980.
- P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* CS-11: 360,393 (Dec. 1963).
- D. Chase, P. A Bello, L-J. Weng, and R. W. Spencer, *Troposcatter Interleaver Study Report*, RADC TR-75-19, Griffiss AFB NY, Feb. 1975.

2704 TURBO CODES

- M. Unkauf and O. A. Tagliaferri, An adaptive matched filter modem for digital troposcatter, *Conf. Rec., Int., Conf. Commun.*, June 1975.
- M. Unkauf and O. A. Tagliaferri, Tactical digital troposcatter systems, Conf. Rec. Nat. Telecomm. Conf., Vol. 2, Dec. 1978, pp. 17.4.1-17.4.5.
- R. W. Lucky, J. Salz, and E. J. Weldon, Jr., Principles of Data Communication, McGraw-Hill, New York, 1986, Chap. 6.
- P. Monsen, Feedback equalization for fading dispersive channels, *IEEE Trans. Inform. Theory* 1T-17: 56-64 (Jan. 1971).
- P. Monsen, Adaptive equalization of the slow fading channel, IEEE Trans. Commun. COM-22: (Aug. 1974).
- 31. P. Monsen, Theoretical and measured performance of a DFE modem on a fading multipath channel, *IEEE Trans. Commun.* **COM-25**(10): (Oct. 1977).
- P. Monsen, Fading channel communications, *IEEE Commun.* Mag. 16-25 (Jan. 1980).
- D. H. Kern and P. Monsen, *MDTS Final Report*, Rep. ECOM-0040-F ECOM, Fort Monmouth, NJ, July 1976.
- 34. J. B. Gadoury, II, Error Performance Characterization Study of a Digital Troposcatter Modem (MD-918/GRC), Defense Communications Agency, Washington, DC, Final Report, July 2, 1983.
- 35. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363–377 (May 1974).
- G. Ungerboeck, Adaptive maximum-likelihood receiver for carrier-modulated data transmission systems, *IEEE Trans. Commun.* COM-22: 624–636 (May 1974).

TURBO CODES

CLAUDE BERROU ALAIN GLAVIEUX ENST Bretagne Brest, France

1. INTRODUCTION

Turbo codes are error-correcting codes that are able to perform in conditions close to the theoretical limits predicted by C. E. Shannon [1]. They were presented to the scientific community in 1993 [2] and at first aroused a certain amount of skepticism, because it was believed at that time that reconciling theory and practice in the matter of channel coding and spectral efficiency would be a longer-term task. In addition, the concepts involved in Turbo coding and decoding, although not revolutionary, were not very familiar to most people in the field. The invention of Turbo codes was the result of a pragmatic construction conducted by C. Berrou and A. Glavieux, based on the intuitions of some European researchers, G. Battail, J. Hagenauer, and P. Hoeher, who in the late 1980s aroused the interest of probabilistic processing in digital communication receivers [3-6]. Previously other researchers, mainly R. Gallager [7] and M. Tanner [8], had already imagined coding and decoding techniques

whose general principles are closely related to those of Turbo codes. Since 1993, Turbo codes have been widely studied, and adopted in several communication systems, and the inherent concepts of the "Turbo" principle have been applied to topics other than error-correcting coding, such as demodulation, detection and multidetection, and equalization.

This article is organized as follows. The next section gives a practical point of view concerning the search for good codes. Sections 3 and 4 are devoted to the building block in Turbo code structure: the Recursive Systematic Convolutional code. Sections 5 and 6 deal with code construction and the permutation issue, and Section 7 presents the Turbo decoding algorithm. Finally, some applications and examples of performance are given.

2. WHAT IS A GOOD ERROR-CORRECTING CODE?

Figure 1 represents, as well as performance without coding, three possible behaviors for an error-correcting coding scheme on an additive white Gaussian noise (AWGN) channel with BPSK or QPSK modulation and rate- $\frac{1}{2}$ coding. To be concrete, the information block is assumed to be around 188 bytes (MPEG application). The error probability P_e of the "no coding" performance is given by the complementary error function, as a function of E_b/N_0 , where E_b is the energy per information bit and N_0 is the monolateral noise density:

$$P_e = \frac{1}{2} \operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right); \operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-u^2) \, du \quad (1)$$

Behavior 1 (in Fig. 1) corresponds to the ideal Shannon system, with the theoretical limit around 0.8 dB, estimated from the report by Dolinar et al. [9]. It assumes random coding and that the minimum distance d_{\min} , deducted from the Gilbert–Varshamov bound [10], would be around 380. The theoretical asymptotic gain, approximated by

$$G_a \approx 10 \log(Rd_{\min})$$
 (2)

would then be higher than 22 dB.

Behavior 2 has good convergence and low d_{\min} . This is, for instance, what is obtained with Turbo coding when the permutation function is not properly designed. Good convergence means that the bit error rate (BER) decreases noticeably, close to the theoretical limit, and low d_{\min} brings about a severe change in the slope, due to an insufficient asymptotic gain. This gain is around 7.5 dB in the example given, and is reached at medium BER ($\approx 10^{-7}$). Below that, the curve remains parallel to the "no coding" one. A possible solution to overcome this *flattening* involves using an outer code, like the Reed–Solomon (RS) code, provided that the statistics of errors is suitable, at the output of the inner decoder, which is not obvious with Turbo codes.

Behavior 3 has poor convergence and high d_{\min} . This is representative of a decoding procedure that does not take advantage of all the information available at the receiver side. A typical example is the classical concatenation of



Figure 1. Possible behaviors for a coding/decoding scheme.

an RS code and a simple convolutional code. Whereas the minimum distance may be very large (but depending on the interleaver depth between the outer and the inner codes), the decoder is clearly suboptimal because the convolutional inner decoder does not take advantage of the RS redundant symbols.

The search for the perfect coding/decoding scheme has always faced the "convergence versus d_{\min} " dilemma. Usually, improving one of either aspect, in some more or less relevant way, weakens the other.

Turbo codes constituted a real breakthrough with regard to the convergence problem. But it was only several years later that, in addition to good convergence properties, sufficient minimum distances were achieved, which then led to powerful standalone channel coding. As can again be observed from Fig. 1, minimum distances as large as those given by random coding are not necessary. Hence, $d_{\min} = 25$ for a 188-byte block with rate $\frac{1}{2}$ would be sufficient to reach the theoretical limit at BER = 10^{-8} , instead of the value 380 given by random coding; $d_{\min} = 32$ would be necessary at BER = 10^{-10} .

To conclude, if a good code has to possess some random properties in order to display the same convergence threshold as random coding, its minimum distance may, in practical cases, be much more reasonable.

3. RECURSIVE SYSTEMATIC CONVOLUTIONAL (RSC) CODES

Pseudorandom generators are widely used in digital communications circuits for encrypting, scrambling, randomizing, spreading, encoding, and other functions. Figure 2 represents a pseudorandom generator, a scrambler, and an RSC encoder, all based on the same linear feedback register (LFR) structure. The register length is denoted ν , also called the *code memory* in the context of coding, and the register state is the vector $\mathbf{S} = (s_1, \ldots, s_j, \ldots, s_{\nu})$. An RSC code is completely defined by two *D* polynomials with degree ν , where *D* is the delay



Figure 2. The same linear feedback register (LFR) structure is used for different fundamental operations, such as pseudorandom generation, scrambling, and encoding.

operator:

$$G_X(D) = 1 + \sum_{j=1}^{\nu-1} G_X^{(j)} D + D^{\nu}$$

$$G_Y(D) = 1 + \sum_{j=1}^{\nu-1} G_Y^{(j)} D + D^{\nu}$$
(3)

where $G_X(D)$ and $G_Y(D)$ are the recursivity and the redundancy polynomials, respectively. $G_X^{(j)}$ (resp. $G_Y^{(j)}$) is equal to 1 if the register tap at level j ($1 \le j \le \nu - 1$) is used in the construction of recursivity (resp. redundancy), and 0 otherwise. $G_X(D)$ and $G_Y(D)$ are generally defined in octal forms. For instance, $1 + D^3 + D^4$ is referenced as polynomial 23. In this section, we consider only the encoding of semi-infinite-length information sequences, beginning at discrete time i = 0 and never ending. In addition to the systematic part of the output: X(D) = d(D), the encoder yields redundancy Y(D), given by

$$Y(D) = \frac{G_Y(D)}{G_X(D)} d(D)$$
(4)

Note that actual redundancy depends not only on the message but also on the LFR initial state, because of the recursivity of the encoder.

Thanks to the linearity property, the code characteristics are expressed with respect to the "all zero" sequence. In this case, any nonzero sequence d(D), accompanied by redundancy Y(D), will represent a possible error pattern for the coding/decoding system. Relation (4) indicates that only a fraction of sequences d(D), which are multiples of $G_X(D)$, lead to finite-length redundancy. We call these particular sequences *return-to-zero* (RTZ) sequences [11], because they force the encoder, if initialized in state 0, to retrieve this state after the encoding of d(D). In what follows, we will be interested only in RTZ patterns, assuming that the decoder will never decide in favor of a sequence whose distance from the "all zero" sequence is infinite. The fraction of sequences d(D) which are RTZ, is exactly

$$p(\text{RTZ}) = 2^{-\nu} \tag{5}$$

because the encoder has 2^{ν} possible states and an RTZ sequence finishes systematically at state 0. Denoting p(NRTZ) as the proportion of non-RTZ sequences (p(NRTZ) = 1 - p(RTZ)), we have

$$\frac{p(\text{NRTZ})}{p(\text{RTZ})} = 2^{\nu} - 1 \tag{6}$$

This is also the maximum possible value for the period L of the pseudorandom generator from which the RSC encoder is derived. This maximum value is obtained if $G_X(D)$ is a prime polynomial.

The shortest RTZ sequence is $G_X(D)$ and any RTZ sequence may be expressed as

$$\operatorname{RTZ}(D) = \sum_{i=0}^{\infty} a_i D^i G_X(D)$$
(7)

where a_i takes value 0 or 1. The minimum number of "1"s belonging to a RTZ sequence is 2. This is because $G_X(D)$ is a polynomial with at least two nonzero terms, and Eq. (5) then guarantees that RTZ(D) also has at least two nonzero terms. In general, the number of "1"s in a particular RTZ sequence is called the *input weight*, or simply the *weight*, and is denoted w. We then have $w_{\min} = 2$ for RSC codes, and the RTZ sequences with weight 2 are of the general form

$$\operatorname{RTZ}_2(D) = D^{\tau}(1 + D^{pL}) \tag{8}$$

where τ is the starting time, p any positive integer, and L the period of the encoder.

RTZ sequences with weight 3 may either exist or not, depending on the expression of $G_X(D)$. For instance, symmetric forms of $G_X(D)$, like polynomial 37, which was used in the first studies on Turbo coding [2], preclude odd values for w. RTZ sequences with even weights always exist, especially of the form

$$\operatorname{RTZ}_{2l}(D) = \sum_{j=1}^{l} D^{t_j} (1 + D^{p_j L})$$
(9)

that is, as a combination of l any weight 2 RTZ sequences. This sort of composite RTZ sequence has to be considered closely when trying to design good permutations for Turbo codes.

RSC codes are decoded with the same trellis approach as classical (nonrecursive, nonsystematic) convolutional codes. The decoding relies either on the Viterbi algorithm [12] for hard-output operation, or the *maximum a posteriori* (MAP) algorithm, also called *a posteriori probability* (APP) or BCJR from the names of its inventors [13], or its simplified versions [14], for softoutput computation, as required by Turbo decoding. The Viterbi algorithm may also be adapted for soft-output decoding purposes [4,5].

Compared with classical convolutional codes, RSC codes offer better performance at low signal-to-noise ratio and/or high coding rates. Figure 3 depicts a significant experiment realized with RSC codes [15]. The MAP algorithm was used to decode RSC codes for different code rates and four increasing values of ν : 2, 4, 6 and 8, and the BER obtained by Monte Carlo simulation was plotted for very low signal-to-noise ratios E_b/N_0 . What is interesting to observe is that, in each case, the four curves corresponding to the different values of v seem to cross at one point (or at worst, in a small cloud of points, probably because the polynomials and the puncturing patterns were chosen arbitrarily). In comparison with the crossing point abscissas, the theoretical limits are indicated by arrows. The conformity between crossing points and theoretical limits suggests that increasing the code memory up to large values, say, several dozens, would offer optimum coding. This is quite natural because the theoretical limits were calculated using the random coding model, and RSC codes become more and more random as the register length increases. When ν tends to infinity, the period $L = 2^{\nu} - 1$ of the maximal-length pseudorandom generator, as well as the ratio between non RTZ and RTZ sequences, as given by (6), also tend to infinity.

We can conclude from this experiment that optimum coding/decoding seems to be achievable by adopting a long RSC code and decoding it using the MAP algorithm. Turbo coding is in fact an artifice for mimicking such a structure while keeping decoding complexity within reasonable boundaries.

In order to achieve coding rates higher than $\frac{1}{2}$, which is the natural rate of the encoder in Fig. 2, "puncturing" may be performed. As many symbols as necessary are discarded before transmission, so as to obtain rates between $\frac{1}{2}$ and 1. In the case of Turbo codes, only redundant symbols from the encoder are punctured. Another possibility for increasing the rate, without (or with less) puncturing, is



Figure 3. Performance of RSC codes for different code memories ($\nu = 2, 4, 6, 8$) and different rates, at very low signal-to-noise ratios, using the MAP decoding algorithm [13]. The Shannon limits, according to Dolinar et al. [9], are indicated by arrows.



Figure 4. The general structure of an *m*-binary RSC encoder with code memory ν . The outputs of the encoder are not represented.

by using m-binary RSC codes, which are at the root of powerful high-rate Turbo codes.

Figure 4 depicts the general structure of an *m*-binary RSC encoder. It uses a pseudorandom generator with code memory ν and generator matrix **G** (size $\nu.\nu$), such that

$$\mathbf{S}_{i+1} = \mathbf{G}\mathbf{S}_i + \mathbf{T}_i \tag{10}$$

where $\mathbf{S}_i = (s_{1,i} \cdots s_{\nu,i})$ is the encoder state vector and $\mathbf{T}_i = (t_{1,i} \cdots t_{\nu,i})$ is the input vector to the encoder taps, both considered at time *i*. The *m*-component input vector $\mathbf{d}_i = (d_{1,i} \cdots d_{m,i})$ is connected to the ν possible taps via a connection grid whose binary matrix, of size $\nu.m$, is denoted **C**. The ν -tap vector \mathbf{T}_i is then given by

$$\mathbf{T}_i = \mathbf{C}\mathbf{d}_i \tag{11}$$

In order to avoid parallel transitions in the corresponding trellis, condition $m \leq \nu$ has to be respected. Except in very particular cases, this encoder is not equivalent to a binary encoder fed successively by d_1, d_2, \ldots, d_m ; that is, the *m*-binary encoder is not generally decomposable.

The redundant output of the machine (not represented in the figure) is calculated, at time i, as

$$y_i = \sum_{j=1\cdots m} d_{j,i} + \mathbf{R}^T \mathbf{S}_i \tag{12}$$

where \mathbf{R}^T is the transposed redundancy vector. The *p*th component of **R** is "1" if the *p*th register tap $(1 \le p \le v)$ is used in the construction of y_i , "0" otherwise. It can easily be shown that y_i can also be written as

$$y_i = \sum_{j=1\cdots m} d_{j,i} + \mathbf{R}^T \mathbf{G}^{-1} \mathbf{S}_{i+1}$$
(13)

provided that

$$\mathbf{R}^T \mathbf{G}^{-1} \mathbf{C} \equiv \mathbf{0} \tag{14}$$

The set of relations (10)-(14) defines completely an *m*-binary RSC code.

On one hand, Eq. (12) ensures that the Hamming weight of $(d_{1,i}, d_{2,i}, \ldots, d_{m,i}, y_i)$ is at least 2, when leaving the reference path (null path), because changing one d value also changes the y value. On the other hand, (13) indicates that the Hamming weight of $(d_{1,i}, d_{2,i}, \ldots, d_{m,i}, y_i)$ is also at least 2 when merging the reference path. Hence, relations (12) and (13) together guarantee that the minimum free distance of the code, whose rate is R = m/(m + 1), is at least 4, whatever m.

As the minimum distance of a concatenated code is much larger than that of its component codes, we can imagine that very large minimum distances may be obtained for Turbo codes, for low as well as high rates. Of course, choosing large values of *m* implies high-complexity decoding because ν also has to be large. For this reason, only values of *m* up to 4 are, for the time being, considered in practice.

4. TERMINATION OF RSC CODES

The optimal decoding of a particular data bit in a convolutionally coded stream requires the knowledge of symbols preceding it and subsequent to it. Therefore, using a convolutional code for encoding a block poses a discontinuity problem at its extremities. With multidimensional codes such as Turbo codes, the question arises each time that the block is encoded. There are three possible solutions, referred to as *trellis termination*, to overcome this problem, which holds for RSC codes as well as for nonrecursive codes:

- 1. Initialize the encoder in state 0 and do nothing about the final state. Data located at the block end are then less protected than the other data. Depending on the target error rate, this penalty may be acceptable. Note, however, that the frame error rate (FER) is more affected than the BER.
- 2. Fix both starting and ending states to 0. The encoder is initialized in state 0 and, after the encoding of the block, the register is forced to state 0 by using ν additional bits, called "tail bits." These tail bits, and the redundant symbols associated with them, are transmitted in order to allow the decoder to choose the most likely path merging to state 0. The actual rate of the code is decreased because of the additional symbols, but the loss is quite negligible for medium and long blocks. This method is not absolutely suitable for multidimensional codes, because the tail bits are encoded only once and the composite code may suffer from this imperfection, which, however, is palpable only at low error rates.
- 3. Use circular (tail-biting) termination. Let us consider an RSC encoder, for instance, the one depicted in Fig. 5 (duobinary code with memory $\nu = 3$). At time i + 1, register state \mathbf{S}_{i+1} is a function of previous state \mathbf{S}_i and tap vector \mathbf{T}_i , as given by (10). For the encoder of Fig. 5, vectors \mathbf{S}_i and \mathbf{T}_i , and matrix \mathbf{G} are given by

$$\mathbf{S}_{i} = \begin{bmatrix} s_{1,i} \\ s_{2,i} \\ s_{3,i} \end{bmatrix}; \quad \mathbf{T}_{i} = \begin{bmatrix} d_{1,i} + d_{2,i} \\ d_{2,i} \\ d_{2,i} \end{bmatrix}; \quad \mathbf{G} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

From Eq. (10) we can infer

$$\mathbf{S}_i = \mathbf{GS}_{i-1} + \mathbf{T}_{i-1}$$

 $\mathbf{S}_{i-1} = \mathbf{GS}_{i-2} + \mathbf{T}_{i-2}$



Figure 5. Recursive convolutional (duobinary) encoder with memory $\nu = 3$. The redundancy output, which is not relevant to the operation of the shift register, has been omitted.

$$\vdots$$

 $\mathbf{S}_1 = \mathbf{G}\mathbf{S}_0 + \mathbf{T}$

Hence, S_i may be expressed as a function of initial state S_0 and of data feeding the encoder between times 0 and *i*:

$$\mathbf{S}_{i} = \mathbf{G}^{i}\mathbf{S}_{0} + \sum_{p=1}^{i}\mathbf{G}^{i-p}\mathbf{T}_{p-1}$$
(15)

If k is the input sequence length (the number of couples for the encoder of Fig. 5), it is possible to find a state \mathbf{S}_c , such that $\mathbf{S}_c = \mathbf{S}_k = \mathbf{S}_0$. Its value is derived from (15)

$$\mathbf{S}_{c} = \left(\mathbf{I} + \mathbf{G}^{k}\right)^{-1} \sum_{p=1}^{k} \mathbf{G}^{k-p} \mathbf{T}_{p-1}$$
(16)

where \mathbf{I} is the $\nu.\nu$ identity matrix. State \mathbf{S}_c depends on the sequence of data and exists only if $\mathbf{I} + \mathbf{G}^k$ is invertible. In particular, k cannot be a multiple of the period L of the recursive generator, which is such that $\mathbf{G}^L = \mathbf{I}$.

The term \mathbf{S}_c is called the *circulation state*. Thus, if the encoder starts from state \mathbf{S}_c , it comes back to the same state when the encoding of the k data (k couples for the encoder of Fig. 5) is completed. Such an encoding process is called *circular* because the associated trellis may be viewed as a circle, without any discontinuity on transitions between states.

Determining \mathbf{S}_c requires a preencoding operation. First, the encoder is initialized in state 0. Then, the data sequence of length k is encoded once, leading to final state \mathbf{S}_k^0 (no redundancy is produced during this operation). Then, from Eq. (15), we obtain

$$\mathbf{S}_k^0 = \sum_{p=1}^k \mathbf{G}^{k-p} \mathbf{T}_{p-1}$$

Combining this result with (16) gives the value of \mathbf{S}_c as follows:

$$\mathbf{S}_{c} = \left(\mathbf{I} + \mathbf{G}^{k}\right)^{-1} \mathbf{S}_{k}^{0} \tag{17}$$

In the second step, data are definitely encoded starting from state \mathbf{S}_{c} .

In practice, the relation between \mathbf{S}_c and \mathbf{S}_k^0 is provided by a small combinational operator with ν input and output bits. The disadvantage of this method lies in having to encode the sequence twice: once from state 0 and the second time from state \mathbf{S}_c . Nevertheless, in most cases, the double-encoding operation can be performed at a frequency much higher than the data rate, so as to reduce the latency effects. Because \mathbf{S}_c is not a priori known by the decoder, the latter has to estimate it by a preliminary step of processing information available preceding \mathbf{S}_c (Fig. 6). So this operation, which we call *prologue*, falls on some data located at the end of the encoded block, and the prologue starts by assigning equal probabilities (or metrics) to all trellis states. The estimate of \mathbf{S}_c is reliable as long as at



Figure 6. As a preamble to normal decoding, a prologue is carried out by the decoder in order to estimate the circulation state \mathbf{S}_c of the circular (tail biting) trellis.

least a dozen or so redundant symbols, are exploited in the prologue.

The circular trellis termination of convolutional codes is a very powerful technique, enabling block encoding for any size and any rate, without the slightest loss in performance. Because a circle has no discontinuity, circular recursive systematic convolutional (CRSC) codes are not weakened by any side effect and, for this reason, are well suited to multidimensional coding.

5. TURBO CODES

A simple means to construct a quasirandom decodable code is a multiple parallel concatenation of CRSC codes, as depicted in Fig. 7. The block of k bits is encoded N

times by *N* CRSC encoders, in a different order each time. Permutations Π_j are drawn at random, except the first one, which is the permutation identity (no permutation). Each component encoder delivers k/N (where k is a multiple of N) redundant symbols, and the global rate is $\frac{1}{2}$. We have already observed (Section 3) that the proportion of RTZ sequences for a simple RSC code, with code memory ν , is $p_1 = 2^{-\nu}$. The proportion of RTZ sequences for the *N*-dimensional code is lowered to

$$p_N = 2^{-N\nu} \tag{18}$$

because the sequence must remain RTZ after *N* different permutations. The other sequences, with proportion $1 - p_N$, yield codewords with a minimum distance at least equal to

$$d_{\min} = \frac{k}{2N} \tag{19}$$

This value assumes that only one sequence is not RTZ (the worst case) and that *Y* redundancy on the corresponding circle takes value "1" every other time statistically. For instance, with N = 8 and $\nu = 3$, we have $p_8 \approx 10^{-7}$, and for sequences of length k = 1024, we obtain $d_{\min} = 64$, which is quite a comfortable minimum distance (see Section 2).

Fortunately, from the complexity standpoint, it is not necessary to adopt such a large dimension. In fact, by replacing random permutation Π_2 with a carefully designed permutation, very good performance can be obtained while limiting the composite code to dimension 2. This is the principle of Turbo coding.



Figure 7. Multiple parallel concatenation of circular recursive systematic convolutional (CRSC) codes. Each encoder delivers k/N redundant symbols, uniformly distributed. Global rate: $\frac{1}{2}$.



Figure 8. Binary and duobinary 8-state Turbo codes with memory $\nu = 3$, using the same RSC encoders (polynomials 13, 15). Natural rates, without puncturing, are $\frac{1}{3}$ and $\frac{1}{2}$, respectively.

Figure 8 represents two Turbo codes, in the classical binary and the duobinary versions. The original message (length *k* bits or couples) is encoded twice, in the natural and the permuted orders, by two RSC codes, denoted C₁ and C₂. In both examples, the component encoders are identical (polynomials 13 for recursivity and 15 for parity redundancy), but this is not a necessity. Natural rates, without puncturing, are $\frac{1}{3}$ and $\frac{1}{2}$. The binary Turbo code is suitable for low code rates ($R \leq \frac{1}{2}$), while the duobinary Turbo code is appropriate for high rates ($R \geq \frac{1}{2}$).

As the permutation function falls on finite-length sequences, the Turbo code is a block code by construction. To distinguish them from concatenated algebraic codes, like product codes, which are decoded by the Turbo algorithm and which were later called *block Turbo codes*, these coding schemes are known as *convolutional Turbo codes* or more technically, as *parallel concatenated convolutional codes* (PCCCs).

The arguments in favor of these coding schemes are as follows:

- 1. The decoding of a convolutionally encoded sequence is very sensitive to errors arriving in packets. Encoding the sequence twice, in different orders, before and after permutation, makes less likely the simultaneous appearance of clustered errors at the decoder inputs of C_1 and C_2 . If packets of errors come to the decoder input of C_1 , the permutation scatters them and they become isolated errors for the decoder of C2, and vice versa. Thus, the bidimensional encoding, formed by either a parallel or a serial concatenation, markedly reduces the vulnerability of convolutional encoding toward packets of errors. But which decoder should one rely on to take the final decision? No criterion allows us to trust either one or the other. The answer is supplied by the Turbo algorithm, which spares us from having to make a choice. This algorithm works out exchanges of probabilistic information between both decoders and forces them to converge toward the same decision, as these exchanges take place.
- 2. Parallel concatenation combines two codes with rates R_1 (code C_1 , with possible puncturing) and R_2 (code C_2 , also with possible puncturing), and the

global rate is

$$R_p = \frac{R_1 R_2}{1 - (1 - R_1)(1 - R_2)} \tag{20}$$

This rate is higher than that of a serially concatenated code $(R_s = R_1R_2)$, for the same values of R_1 and R_2 , and the lower these rates, the larger the difference. Thus, with the same performance of component codes, parallel concatenation offers a better global rate, but this advantage is lost when the rates come close to unity.

3. Parallel concatenation relies on systematic codes. At least one of these codes has to be recursive, for a fundamental reason related to the minimum data input weight (w_{\min}) . Figure 9 depicts two nonrecursive systematic convolutional codes, concatenated in parallel. The input sequence is "all zero," except in one position. This single "1" disrupts the encoder output during a short lapse of time, given by the constraint length. Actually, this sequence with only one "1" is the minimal RTZ sequence of any nonrecursive code. Thus, redundancy Y_1 is very poor with respect to this particular sequence. After the permutation, the sequence remains "all zero" except in one position, and redundancy Y_2 is as poor as the first one. In fact, the minimum distance of this



Figure 9. The parallel concatenation of nonrecursive systematic convolutional codes makes up a poor code with regard to weight 1 sequences.

composite code is not higher than that obtained from a single code, with the same rate. If we replace at least one of the nonrecursive encoders by a recursive one, the considered input sequence is no longer an RTZ sequence for this encoder, and the redundancy weight is considerably increased.

4. As seen at the beginning of this section, it is possible to increase the Turbo code dimension N by using more than two component encoders. The result is a noticeable increase in the minimum distance; with $N \ge 5$ and a set of random permutations, the Turbo code is comparable to a quasirandom code. Unfortunately, the convergence threshold of the decoder (the signal-to-noise ratio above which most errors are corrected) deteriorates when the dimension is increased. This is due to the Turbo decoding principle, which is to consider iteratively each dimension, one after the other. As the redundancy rate of each component code decreases when their number increases, the first steps of the decoding are penalized in comparison with the two-dimensional code. This conflict between large minimum distance and low convergence threshold, already mentioned in Section 2, is a permanent feature of error correction coding.

A particular Turbo code is defined by

- m, the number of bits in the input words. Applications known so far consider binary (m = 1) and duobinary (m = 2) words. More recent decoding techniques using the dual-code approach [16] could allow larger values of m to be considered without increasing the decoding complexity too much.
- The component codes C_1 and C_2 (code memory ν , recursivity and redundancy polynomials). The values of ν are 3 or 4 in practice and the polynomials are generally those that are recognized as the best for simple unidimensional convolutional coding, that is, (15,13) for $\nu = 3$ and (23,35) for $\nu = 4$, or their symmetric forms.
- The permutation function, which plays a decisive role when the target BER is lower than about 10^{-5} . Above this value, the permutation may be any, provided obviously that it respects at least the scattering property (e.g., the permutation may be the regular one).
- The puncturing pattern. This has to be as regular as possible, such as that for simple convolutional codes.

In addition to this rule, the puncturing pattern is defined in close relationship with the permutation function when very low errors rates are sought. Puncturing is achieved on the systematic part of the codewords. In some cases, it may be conceivable to puncture the redundant part instead, with the aim of increasing the minimum distance. This is then achieved to the detriment of the convergence threshold, because, from this standpoint, deleting data shared by all the decoders is more penalizing than deleting data that are useful to only one of them.

6. THE PERMUTATION FUNCTION

Either called *permutation* or *interleaving*, the technique involving scattering data over time is of great service in digital communications. It is used, for instance, to reduce the effects of dimming in fading channels, and more generally to combat perturbations that affect consecutive symbols. In the case of Turbo codes, permutation also plays this role for at least one dimension of the composite code. But its importance goes beyond this; permutation also fixes the minimum distance of the concatenated code, in close relationship to the properties of component codes.

Let us consider the binary Turbo code represented in Fig. 8a, with permutation falling on k bits. The worst permutation we can imagine is permutation identity, which minimizes the coding diversity $(Y_1 = Y_2)$. On the other hand, the best permutation that could be used, but that probably does not exist [17], could allow the concatenated code to be equivalent to a sequential machine whose irreducible number of states would be 2^{k+6} . There are actually k + 6 binary storage elements in the structure: k in the permutation memory and 6 in the encoders. Assimilating this machine to a convolutional code would give a very long code and very large minimum distances, for usual values of k. From the worst to the best of permutations, there is great choice between the k! possible combinations, and we still lack a sound theory about it. Nevertheless, good permutations have already been designed to elaborate normalized Turbo codes, using pragmatic approaches.

6.1. Regular Permutation

The starting point in the design of a permutation is the regular permutation, illustrated in two ways in Fig. 10, for binary codes. The first one assumes that the block of k bits can be organized as a table with M rows and N columns (k = M.N). The permutation then involves writing data



Figure 10. Regular permutation with rectangular (**a**) or circular (**b**) forms.

2712 TURBO CODES

linewise in an appropriate memory and reading them columnwise, possibly with skips of columns [18]. The second one is used without any hypothesis on the value of k. After writing the data in a linear memory, with address $i \ (0 \le i \le k - 1)$, the block is likened to a circle, and both extremities of the block (i = 0 and i = k - 1) then become contiguous. The data are read out such that the *j*th datum read was written at position *i* given by

$$i = Pj \mod k \tag{21}$$

where *P* is an integer, prime with *k*. In order to maximize the spatial distance after permutation between two consecutive data, whatever they are, and vice versa, *P* has to be close to $\sqrt{2k}$, with the condition

$$k \approx \frac{P}{2} \mod P$$
 (22)

6.2. The Statistical Approach

An upper bound of the error probability P_e of a code, assuming optimal decoding, is given by

$$P_e \le \sum_{d=d_{\min}}^{\infty} M_d \operatorname{erfc}\left(\sqrt{Rd\frac{E_b}{N_0}}\right)$$
 (23)

where d_{\min} is the minimum distance, R the rate considered, and M_d the sum of the weights, called *multiplicity*, of all codewords at distance d. When designing a permutation, the aim is thus to maximize d_{\min} and minimize the multiplicities. Before doing this work, it is interesting to have an idea about the performance that any typical permutation might give. Benedetto and Montorsi [19] proposed to use uniform (or statistical) model of permutation, which is a device that associates with a message of length k and weight w, one of the possible $\binom{k}{w}$ messages obtained by the permutation of w bits among k. The $\binom{k}{w}$ permuted messages have the same probability $\frac{1}{\binom{k}{w}}$ of being present at the second

encoder input.

This statistical interleaver performs similarly to an interleaver that would be the average of all possible deterministic interleavers with size k. Therefore, there exists at least one interleaver with fixed rules that allows one to reach, and even exceed, the performance given by the uniform interleaver. In fact, it is easy to find deterministic permutations better than the statistical one when the target error rate is low.

Let us assume that A_{wd} denotes the number of codewords at distance d and with weight w, without the permutation. It is then demonstrated [19] that the uniform permutation modifies this value into $w!k^{1-w}A_{wd}$, so the value is reduced if $w!k^{1-w}$ is less than 1. The worst case is given by the minimum value of w, that is, w_{\min} , which is 1 for algebraic codes (BCH, etc.) and nonrecursive convolutional codes, but 2 for RSC codes. The term $k^{1-w_{\min}}$,

called *interleaving gain*, is then favorable only in the latter case and is equal to $k^{1-2} = 1/k$. It also follows from this result that the longer the encoded block, the lower the multiplicities and the better the performance. This point is in agreement with the theoretical limits obtained on finite-length blocks [9].

In conclusion, the statistical approach of the permutation problem confirms the need to use RSC codes and gives a means of observing and estimating an interleaving gain that depends on the block size. This gain is essentially visible at large error rates (BER $\geq 10^{-5}$). For lower error rates, the main parameter is the minimum distance, which has to be maximized, and the uniform interleaver is not suitable for this.

6.3. Real Permutations

The dilemma in the design of a good permutation lies in the need to obtain a sufficient minimum distance for two distinct classes of codewords, which require conflicting treatment [20]. The first class contains all codewords with input weight $w \leq 3$, and a good permutation for this class is as regular as possible. The second class encompasses all codewords with input weight w > 3, and nonuniformity (controlled disorder) has to be introduced in the permutation function to obtain a large minimum distance. Figure 11 illustrates the situation, showing the example of a rate- $\frac{1}{3}$ Turbo code, using component binary encoders with code memory v = 3 and periodicity $L = 2^v - 1 = 7$.

For the sake of simplicity, the block of k bits is organized as a rectangle with M rows and N columns ($M \approx N \approx \sqrt{k}$). Regular permutation is used; that is, data are written linewise and read columnwise:



Figure 11. Some possible RTZ (return-to-zero) sequences for both encoders C_1 and C_2 , with $G_X(D) = 1 + D + D^3$ (period L = 7): (a) with input weight w = 2; (b) with w = 3; (c) with w = 6 or 9.

- Case (a) in Fig. 11 depicts a situation where encoder C_1 (the horizontal one) is fed by an RTZ sequence with input weight w = 2. Redundancy Y_1 delivered by this encoder is poor, but redundancy Y_2 produced by encoder C_2 (the vertical one) is very informative for this pattern, which is also an RTZ sequence, but whose span is 7.*M* instead of 7. The associated minimum distance would be around 7.M/2, which is a large minimum distance for typical values of *k*. With respect to this w = 2 case, the code is said to be "good" because d_{\min} tends to infinity when *k* tends to infinity.
- Case (b) in Fig. 11 deals with a weight 3 RTZ sequence. Again, whereas the contribution of redundancy Y_1 is not high for this pattern, redundancy Y_2 gives relevant information over a large span, of length 3.M. The conclusions are the same as for case (a).
- Case (c) in Fig. 11 shows two examples of sequences with weights w = 6 and w = 9, which are RTZ sequences for both encoders C₁ and C₂. They are obtained by a combination of two or three minimallength RTZ sequences. The set of redundancies is limited and depends on neither *M* nor *N*. These patterns are typical of codewords that limit the minimum distance of a Turbo code, when using a regular permutation.

In order to "break" rectangular patterns, some disorder has to be introduced into the permutation rule, while ensuring that the good properties of regular permutation, with respect to weights 2 and 3, are not lost. This is the crucial problem in the search for good permutation, which has not yet found a definitive answer. Nevertheless, some good permutations have already been devised for several applications (CCSDS [21], IMT-2000 [22,23], DVB [24,25]).

7. TURBO DECODING

Decoding a composite code by a global approach is not possible in practice because of the tremendous number of states to consider. A joint probabilistic process by the decoders of C_1 and C_2 has to be elaborated. Because of latency constraints, this joint process is worked out in an iterative manner in a digital circuit (analog versions of the Turbo decoder are also considered, offering much larger throughputs [26]).

Turbo decoding relies on the following fundamental criterion, which is applicable to all "message passing" or "belief propagation" [27] algorithms:

When having several probabilistic machines work together on the estimation of a common set of symbols, all the machines have to give the same decision, with the same probability, about each symbol, as a single (global) decoder would.

To make the composite decoder satisfy this criterion, the structure of Fig. 12 is adopted. The double loop enables both component decoders to benefit from the whole redundancy. The term "Turbo" was given to this feedback construction with reference to the principle of the turbo-charged engine.

The components are soft-in/soft-out (SISO) decoders, and permutation (Π) and inverse permutation (Π^{-1}) memories. The node variables of the decoder are logarithms of likelihood ratios (LLRs). An LLR related to a particular binary datum d_i is defined as

$$LLR(d_i) = \ln\left(\frac{\Pr(d_i = 1)}{\Pr(d_i = 0)}\right)$$
(24a)

For a decoder processing *m*-binary words instead of binary data, the LLRs associated with the 2^m possible values of the word vector \mathbf{d}_i , could be written as

$$LLR_{j}(\mathbf{d}_{i}) = \ln\left(\frac{\Pr(\mathbf{d}_{i} \equiv j)}{\Pr(\mathbf{d}_{i} \equiv 0)}\right)$$
(24b)

where $Pr(\mathbf{d}_i \equiv j)$ is the probability that vector \mathbf{d}_i takes the *j*th value, numbered from 0 to $2^m - 1$. Because LLR₀ is always equal to 0, there are only $2^m - 1$ LLRs to calculate in practice.

The role of a SISO decoder is to process an input LLR and, thanks to local redundancy (i.e., y_1 for DEC1, y_2 for



Figure 12. An 8-state Turbo code and its associated decoder (basic structure assuming no delay processing).

DEC2), to try to improve it. The output LLR of a SISO decoder, for a binary datum, may be simply written as

$$LLR_{out}(d) = LLR_{in}(d) + z(d)$$
(25)

where z(d) is the *extrinsic* information about d, provided by the decoder. If this works properly, z(d) is usually negative if d = 0, and positive if d = 1.

The composite decoder is constructed in such a way that only extrinsic terms are passed by one component decoder to the other. The input LLR to a particular decoder is formed by the sum of two terms: the information symbols (x) stemming from the channel and the extrinsic term (z)provided by the other decoder, which serves as a priori information. The information symbols are common inputs to both decoders, which is why the extrinsic information must not contain them. In addition, the outgoing extrinsic information does not include the incoming extrinsic information, in order to reduce correlation effects in the loop.

The practical course of operation is

- Step 1. Process the data peculiar to one code, say, C_2 (x and y_2) by decoder DEC2, and store the extrinsic pieces of information (z_2) resulting from the decoding in a memory. If data are missing because of puncturing, the corresponding values are set to analog 0 (neutral value).
- Step 2. Process the data specific to C_1 (x, deinterleaved z_2 and y_1) by decoder DEC1, and store the extrinsic pieces of information (z_1) in a memory. By properly organizing the read/write instructions, the same memory can be used for storing both z_1 and z_2 .
- Steps 1 and 2 make up the first iteration.
- Step 3. Process C_2 again, now taking interleaved z_1 into account, and store the updated values of z_2 .

And so on.

The process ends after a preestablished number of iterations, or after the decoded block has been estimated as correct, according to some stop criterion (see the report by Matache et al. [28] for possible methods of stopping rules). The typical number of iterations for the decoding of convolutional Turbo codes is 4-10, depending on the constraints relating to complexity, power consumption, and latency.

According to the structure of the decoder, after p iterations, the output of DEC1 is

$$LLR_{out1,p}(d) = (x + z_{2,p-1}(d)) + z_{1,p}(d)$$

where $z_{u,p}(d)$ is the extrinsic piece of information about d, yielded by decoder u after iteration p, and the output of DEC2 is

LLR_{out2,p}(d) =
$$(x + z_{1,p-1}(d)) + z_{2,p}(d)$$

If the iterative process converges toward fixed points, $z_{1,p}(d) - z_{1,p-1}(d)$ and $z_{2,p}(d) - z_{2,p-1}(d)$ both tend to zero when p goes to infinity. Therefore, from the equations above, both LLRs become equal, which fulfills the

fundamental condition of equal probabilities provided by the component decoders for each datum d. As for the proof of convergence itself, one can refer to various papers dealing with the theoretical aspects of the subject [e.g., 29,30].

Turbo decoding is not optimal. This is because an iterative process obviously must begin, during the first half-iteration, with only a part of the redundant information available (either y_1 or y_2). Fortunately, loss due to suboptimality is small: about 0.5 dB for binary Turbo codes and 0.3 dB for duobinary Turbo codes.

There are two families of SISO algorithms: those based on the Viterbi algorithm [12], which can be used for high throughput continuous stream applications; and others based on the MAP (also known as BCJR or APP) algorithm [13] or its simplified derived versions [14], for block decoding. If the full MAP algorithm is chosen, it is better for extrinsic information to be expressed by probabilities instead of LLRs, which avoids the need to calculate a useless variance for extrinsic terms.

The following practical parameters must be factored into the design of a Turbo decoder that processes LLRs:

- The number of quantization bits for the channel samples: typically 3 or 4 if the code is associated with BPSK or QPSK modulation; 5 or 6 with higher-order modulations that require greater accuracy.
- The number of quantization bits for extrinsic information: typically 1 bit more than those of channel samples.
- The scale factor, which is the ratio between the mean absolute value of the channel data and its maximum absolute value. This factor depends on the coding rate, on the type of channel and also on quantization accuracy.

In practice, depending on the kind of SISO algorithm chosen, some tuning operations (multiplying, limiting) on extrinsic information are added to the basic structure to ensure stability and convergence within a small number of iterations.

8. APPLICATIONS

Table 1 summarizes normalized applications of convolutional Turbo codes known to date. These use either 8-state binary or duobinary RSC component encoders or 16-state binary encoders. Figure 13 shows some representative examples of performance obtained from various Turbo codes associated with QPSK and 8-PSK modulation on Gaussian channels. For the latter case, the so-called pragmatic scheme [31], which is the simplest way to combine channel coding and modulation, was used. The component decoding algorithm was the max-log-MAP (also called subMAP) algorithm [14], which is derived from the exact MAP procedure [13] by doing operations in the logarithmic domain (multiplications become additions), and by replacing additions by maximum (Max) functions.

To simplify, let us say that 8-state Turbo codes are suitable for the medium error rates (FER $\approx 10^{-4}$) that are

Application	Turbo Code	Termination	Polynomials	Rates
CCSDS [21]	Binary, 16-state	Tail bits	23, 33, 25, 37	$\frac{1}{6}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$
IMT-2000 [22,23]	Binary, 8-state	Tail bits	15, 13, 17	$\frac{1}{4}, \frac{1}{3}, \frac{1}{2}$
DVB-RCS [24]	Duobinary, 8-state	Circular	15, 13	$\frac{1}{3} - \frac{6}{7}$
DVB-RCT [25]	Duobinary, 8-state	Circular	15, 13	$\frac{1}{2}, \frac{3}{4}$
Inmarsat (mini M)	Binary, 16-state	No	23, 35	$\frac{1}{2}$
Eutelsat (skyplex)	Duobinary, 8-state	Circular	15, 13	$\frac{4}{5}, \frac{6}{7}$

Table 1. Applications of Convolutional Turbo Codes



Figure 13. Some examples of performance, expressed in FER, achievable with Turbo codes on Gaussian channels. In all cases, decoding was performed using the max-log-MAP algorithm [14] with eight iterations and 4-bit input quantization.

required, for instance, by ARQ (Automatic Repeat reQuest) systems, whereas 16-state Turbo codes are necessary when the target error rates are lower (FER $\approx 10^{-8}$), for broadcasting in particular.

Acknowledgment

Warm thanks to Janet Ormrod for her contribution to the writing of this article.

BIOGRAPHIES

Claude Berrou was born in Penmarc'h, France, in 1951. He received the Electrical Engineering Degree from the Institut National Polytechique, Grenoble, France, in 1975. In 1978, he joined the Ecole Nationale Supérieure des Télécommunications de Bretagne, Brest, France, where he is currently Professor in the Electronics Department. His research topics include algorithm-silicon interaction, electronics and digital communications, error-correcting codes, Turbo codes that he discovered in 1991, softin/soft-out decoders, and genetic coding. He is the author or co-author of eight registered patents and about 40 publications in the field of Turbo coding. He was the corecipient of the 1997 IEEE Trans. com. Paper Award and the 1998 IEEE Information Theory Society Paper Award. He also received the Médaille Ampère (SEE) and one of the Golden Jubilee Awards for Technological Innovation (IEEE IT Society) in 1998.

Alain Glavieux was born in Paris, France, in 1949. He received the Electrical Engineering Degree from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1978. In 1979, he joined the Ecole Nationale Supérieure des Télécommunications de Bretagne, Brest, France, where he is currently Professor and Director of Corporate Relations. His research interest includes Turbo coding, Turbo equalization, and communications over fading channels. He was the co-recipient of the 1997 IEEE Trans. Com. Paper Award and the 1998 IEEE Information Theory Society Paper Award. He also received one of the Golden Jubilee Awards for Technological Innovation (IEEE IT Society) in 1998.

BIBLIOGRAPHY

- C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27: (July, Oct. 1948).
- 2. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: turbo-codes, *Proc. IEEE ICC '93*, Geneva, May 1993, pp. 1064–1070.
- G. Battail, Coding for the Gaussian channel: The promise of weighted-output decoding, Int. J. Satellite Commun. 7: 183-192 (1989).
- G. Battail, Pondération des symboles décodés par l'algorithme de Viterbi, Ann. Télécommun. 42(1–2): 31–38 (Jan. 1987).
- J. Hagenauer and P. Hoeher, A Viterbi algorithm with softdecision outputs and its applications, *Proc. GLOBECOM* '89, Dallas, TX, Nov. 1989, pp. 47.11-47.17.
- J. Hagenauer and P. Hoeher, Concatenated Viterbi-decoding, Proc. Int. Workshop on Information Theory, Gotland, Sweden, Aug.-Sept. 1989.
- R. G. Gallager, Low-density parity-check codes, *IRE Trans.* Inform. Theory IT-8: 21–28 (Jan. 1962).
- R. M. Tanner, A recursive approach to low complexity codes, IEEE Trans. Inform. Theory IT-27: 533-547 (Sept. 1981).
- S. Dolinar, D. Divsalar, and F. Pollara, Code Performance as a Function of Block Size, TMO Progress Report 42-133, JPL, NASA, May 1998.
- 10. J. G. Proakis, *Digital Communications*, 2nd ed., McGraw-Hill, New York, 1989, pp. 426-428.

2716 TURBO EQUALIZATION

- R. Podemski, W. Holubowicz, C. Berrou, and G. Battail, Hamming distance spectra of turbo-codes, *Ann. Télécommun.* 50(9-10): 790-797 (Sept.-Oct. 1995).
- G. D. Forney, The Viterbi algorithm, Proc. IEEE 61(3): 268–278 (March 1973).
- L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 248–287 (March 1974).
- P. Robertson, P. Hoeher, and E. Villebrun, Optimal and suboptimal maximum a posteriori algorithms suitable for turbo decoding, *Eur. Trans. Telecommun.* 8: 119-125 (March-April 1997).
- C. Berrou, Some clinical aspects of turbo codes, Proc. 1st Int. Symp. Turbo Codes & Related Topics, Brest, France, Sept. 1997, pp. 26-31.
- J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* 42(2): 429–445 (March 1996).
- Y. V. Svirid, Weight distributions and bounds for turbo-codes, Eur. Trans. Telecommun. 6(5): 543–555 (Sept.–Oct. 1995).
- E. Dunscombe and F. C. Piper, Optimal interleaving scheme for convolutional coding, *Electron. Lett.* 25(22): 1517–1518 (Oct. 1989).
- S. Benedetto and G. Montorsi, Design of parallel concatenated convolutional codes, *IEEE Trans. Commun.* 44(5): 591-600 (May 1996).
- C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* 44(10): 1261-1271 (Oct. 1996).
- 21. Consultative Committee for Space Data Systems, Recommendations for Space Data Systems. Telemetry Channel Coding, Blue Book, May 1998.
- 22. 3GPP Technical Specification Group, *Multiplexing and Channel Coding (FDD)*, TS 25.212 v2.0.0, June 1999.
- 23. TIA/EIA/IS 2000-2, Physical Layer Standard for cdma2000 Spread Spectrum Systems, July 1999.
- DVB, Interaction Channel for Satellite Distribution Systems, ETSI EN 301 790, V1.2.2, Dec. 2000, pp. 21–24.
- 25. DVB, Interaction Channel for Digital Terrestrial Television, ETSI EN 301 958, V1.1.1, Aug. 2001, pp. 28–30.
- H.-A. Loeliger, F. Tarkoy, F. Lustenberger, and M. Helfenstein, Decoding in analog VLSI, *IEEE Commun. Mag.* 37(4): 99–101 (April 1999).
- R. J. McEliece and D. J. C. MacKay, Turbo decoding as an instance of Pearl's "belief propagation" algorithm, *IEEE J. Select. Areas Commun.* 16(2): 140–152 (Feb. 1998).
- A. Matache, S. Dolinar, and F. Pollara, Stopping Rules for Turbo Decoders, TMO Progress Report 42-142, JPL, NASA, Aug. 2000.
- 29. Y. Weiss and W. T. Freeman, On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs, *IEEE Trans. Inform. Theory* **47**(2): 736–744 (Feb. 2001).
- L. Duan and B. Rimoldi, The iterative turbo decoding algorithm has fixed points, *IEEE Trans. Inform. Theory* 47(7): 2993–2995 (Nov. 2001).
- S. Le Goff, A. Glavieux, and C. Berrou, Turbo-codes and high spectral efficiency modulation, *Proc. IEEE ICC'94*, New Orleans, May 1994, pp. 645–649.

TURBO EQUALIZATION

KRISHNA R. NARAYANAN Texas A&M University College Station, Texas

1. INTRODUCTION

Intersymbol interference (ISI) can often be a limiting factor when communicating through band-limited channels and, hence, it is important to use efficient equalization techniques to combat the effect of ISI. A good discussion of equalization techniques for uncoded systems including both sequence estimation type equalizers and symbol-bysymbol equalizers can be found in another study [1]. When an error correction code (ECC) is used in conjunction with an ISI channel, the equalizer should take advantage of the error correction capability of the code. The optimal receiver which performs joint equalization and decoding can be computationally complex and is usually not implementable in practice. Therefore, several suboptimal solutions have been studied. One straightforward technique is to perform equalization without taking the code structure in to account, followed by decoding. Since the operating signalto-noise ratio (SNR) for coded systems is usually much smaller than that for the uncoded case, the performance of the equalizer can be severely affected. Techniques that use the tentative decision from the decoder in the equalization step, such as in delayed decision feedback sequence estimation [2], have been shown to perform better than separate equalization and decoding.

In 1995, Douillard et al. proposed another sub-optimal joint equalization and decoding technique [3] by extending the idea of iterative decoding that was used to decode Turbo codes [4]. The idea was to exchange soft information between a soft-input soft-output (SISO) equalizer and an SISO decoder in an iterative fashion and they naturally named it Turbo equalization. Since then, several researchers have shown that turbo equalization can significantly improve the performance over separate equalization and decoding, and is a practical solution to obtaining close to capacity performance on ISI channels. Currently, Turbo equalization is being considered for use in future-generation digital magnetic recording systems and wireless systems. Here, we will explain the main concepts behind turbo equalization and summarize some of the results in this area.

2. SYSTEM MODEL

A typical system model with an error correction code and an ISI channel is shown in Fig. 1. A block of K bits of the binary data sequence a is first encoded by an ECC (referred to as *outer code* here) into N coded bits. Any code that permits efficient SISO decoding can be used, but we will restrict our attention to a convolutional code, parallel concatenated convolutional code (PCCC or turbo code), or low-density parity-check (LDPC) code as the outer code. The multiplexed output x is interleaved and the interleaved sequence y is modulated into z. We will assume



Figure 1. Discrete-time system model.

that the modulation is memoryless and the modulated sequence z is then transmitted over the ISI channel. At the receiver, the received signal is passed through a whitened matched filter (WMF) and the output of the WMF is sampled every T seconds, where T is the symbol duration. The combination of the ISI channel, WMF, and the sampler is equivalent to a discrete-time transversal filter (DTTF). The DTTF can be represented by an L tap-delay line with weights $f_{-l_1}, \ldots, f_{-1}, f_0, f_1, \ldots, f_{l_2}$, as shown in Fig. 1, where l_1 and l_2 are the number of postcursor and precursor taps, respectively, and $L = l_1 + l_2 + 1$. Therefore, the WMF output at time instant k can be expressed as

$$r_k = \sum_{i=-l_1}^{l_2} z_{k-i} \quad f_i + n_k,$$
(1)

where n_k are samples of a white Gaussian noise process with zero mean and variance σ_{ch}^2 .

3. TURBO EQUALIZATION ALGORITHM

We will start with a few preliminaries and then describe the Turbo equalization algorithm. To keep the discussion simple and clear, let us assume that a is a binary sequence of equiprobable and independent bits, the outer code is a binary convolutional code, and that the modulation is binary phase shift keying (BPSK). Further, let us assume that the receiver has perfect knowledge of the tap coefficients and the variance of the additive noise. For any sequence x, let x_k^N denote the sequence x from time kto N. The conditional log-likelihood ratio (LLR) of a binary random variable $x_k \in \{0, 1\}$ given a noisy observation of x, namely r, is defined as

$$\Lambda(x_k) \stackrel{\scriptscriptstyle \Delta}{=} \log \frac{P(x_k = 1|r)}{P(x_k = 0|r)} \tag{2}$$

The optimal receiver that minimizes the bit error rate (BER) computes an estimate \hat{a}_k according to

$$\hat{a}_{k} = \frac{1}{0}, \quad \text{if } P(a_{k} = 1|r_{1}^{N}) > P(a_{k} = 0|r_{1}^{N}) \\ 0, \quad \text{otherwise}$$
(3)

Although conceptually simple, it is quite difficult (almost impossible) to implement the above receiver for most cases of practical interest. The Turbo equalization algorithm computes a suboptimal solution to the problem presented above as explained below. We will first give the steps of the Turbo equalization algorithm without explaining the details of the computation at each step. In the next sections, we will detail the computations to be performed at each step. The overall algorithm is an iterative algorithm that consists of an SISO equalizer and an SISO decoder that exchanges LLR estimates of the bits x_k in the sequence x. At the *m*th stage, the equalizer provides extrinsic information in the form of a loglikelihood ratio on each bit x_k , denoted by $L_i^m(x_k)$. The extrinsic information generated by the equalizer $L_i^m(x_k)$ does not include any information on x_k given by the decoder. It represents the contribution of the received signal r and the *a priori* information on x_i given by the outer decoder for all $i \neq k$. In the following section, we will use L to denote extrinsic LLRs and Λ to denote the total LLR. Subscripts *i* and *o* denote quantities generated from the inner SISO (equalizer) and outer SISO, respectively. Superscript mrefers to the iteration number; and when there is no danger of confusion, we will drop the superscript. The outer decoder uses the extrinsic information $L_i^m(x_k)$ as though they were the output of a hypothetical channel and provides extrinsic information on x_k , denoted by $L_{o}^{m}(x_{k})$. This extrinsic information $L_{o}^{m}(x_{k})$ is based purely on the code constraints imposed by the outer code, and is used as a priori information in the (m+1)th stage in the equalizer (inner decoder). The outer decoder also produces likelihood ratios for the information bits $\Lambda^m(a_k)$ from which hard-decision estimates \hat{a}_k can be obtained. The iterations proceed until a stopping criterion (e.g., a cyclic redundancy check) is satisfied or a maximum of Miterations have been performed. Note that the extrinsic information generated by the outer decoder should be interleaved before being used in the equalizer and the extrinsic information generated by the equalizer should be deinterleaved at each stage as shown in Fig. 2. The turbo equalization algorithm can be summarized as follows:

- 1. Initialization: $L_o^0(x_k) = 0, \forall k$.
- 2. Iterations: During the *m*th iteration, for m = 1, 2, ..., M
 - a. *SISO equalization*: compute extrinsic LLR for the coded bits in the sequence *x* based on the extrinsic information provided by the outer decoder in the previous iteration and the received signal *r*. Formally, we can denote the output of the SISO equalizer as

For k = 1, 2, ..., N, compute $L_i^m(x_k)$ based on r, and

$$(L_o^{m-1}(x_1), \dots, L_o^{m-1}(x_{k-1}),$$

 $L_o^{m-1}(x_{k+1}), \dots, L_o^{m-1}(x_N))$



Figure 2. Turbo equalizer structure.

b. SISO decoding: compute extrinsic LLRs for the coded bits in the sequence x and the total LLRs for the data bits in the sequence a based on the extrinsic information provided by the SISO equalizer (generated in the previous step):

For k = 1, 2, ..., N, compute $L_o^m(x_k)$ based on $(L_i^m(x_1), ..., L_i^m(x_N))$

For k = 1, 2, ..., K, compute $\Lambda^m(a_k)$ based on $(L_i^m(x_1), ..., L_i^m(x_N))$

- c. Hard decision $\hat{a}_k^m = 1$ if $\Lambda(a_k) > 0$ and $\hat{a}_k^m = 0$ otherwise.
- d. Check for stopping criterion and if not satisfied, set $m \leftarrow m + 1$ and go to step (a).

4. SOFT-OUTPUT EQUALIZATION

Several algorithms are used to implement the SISO equalizer providing an option to tradeoff complexity for performance. These can be broadly classified into trellis based approaches and filtering based approaches as described below.

4.1. Trellis-Based Approaches

These algorithms exploit the trellis structure of the ISI channel or, equivalently, the Markov nature of the outputs of the ISI channel to efficiently compute the extrinsic LLRs (or APPs). We first note that the ISI channel with L taps can be considered as a convolutional code with constraint length L and a trellis structure can be associated with the channel. At each time instant k, the input to the channel z_k and the state of the ISI channel (which corresponds to the previous L - 1 bits) determine the next state. Along with the channel tap coefficients they also determine the output at each time instant. The trellis structure of a 2-tap ISI channel with taps f_0 and f_1 is shown in Fig. 3.



Figure 3. 2-tap ISI channel and trellis.

The optimal equalizer given the channel impulse response and the noise variance at the receiver can be implemented using the Bahl, Cocke, Jelinek, and Raviv (BCJR) algorithm [5]. The BCJR algorithm produces optimal a posteriori probabilities (APP's) on x_k given the received signal *r* by using forward and backward recursions on the trellis of the ISI channel. For an explanation of the algorithm and implementation aspects, the reader is referred to Ref. 6. The BCJR algorithm is optimal; however, its complexity is quite high. Lowercomplexity variants of the BCJR algorithm such as the max-log-MAP can also be used [7]. Another alternative is the use of the soft-output Viterbi algorithm (SOVA) as proposed by Hoeher and Hagenauer [8], which modifies the hard-output Viterbi algorithm (VA) to provide reliabilities for the decoded bits in addition to the hard decisions. The SOVA is less complex compared to the BCJR algorithm; however, it is suboptimal. It is also known that the soft-output of the SOVA is optimistic [9] and that the performance can be improved by scaling the extrinsic information produced by the SOVA. Douillard et al. use the SOVA with scaling in their first paper on Turbo equalization [3]. The decoding complexity of the BCJR algorithm, max-log MAP algorithm, and the SOVA increases exponentially with L and even for moderate L, it becomes difficult to implement these algorithms.

One way to reduce the complexity is to use an SISO algorithm based on reduced-state sequence estimation such as the M and T BCJR algorithms [10]. In M-type algorithms, only the best M paths in each stage of the trellis are retained and the rest of the paths are discarded. Hence, the complexity is independent of the channel memory and depends only on *M*. In *T*-type algorithms, all paths whose metric is less than a threshold are dropped and, hence, complexity reduction is achieved. However, the reduction in complexity depends on the channel conditions. When used in turbo equalization, during the first few iterations, more number of paths will be retained and as the iterations progress, the effective channel has a higher SNR and, hence, the complexity dynamically decreases. The concepts of M and T algorithms have now been extended to the SOVA and shown to perform well for the case of Turbo equalization [11]. Another technique to reduce the equalization complexity is to truncate the channel memory or use hard-decision feedback from the decoder to cancel the ISI due to some of the taps and use a trellis based equalizer with the shortened channel [12].

4.2. Soft Interference Cancellation with Filtering Approach

A different approach to soft output equalization is to modify conventional hard-output decision feedback equalizers in order to accept and provide soft output. The main advantage of this approach is that the complexity is only $O(L^2)$ compared to the exponential dependence on L for the optimal equalizer. Several different forms of this equalizer have been proposed [13–15]. We will now describe the one due to Wang and Poor [14] in detail. We begin by rewriting (1) in matrix form:

$$\mathbf{F}_{\mathbf{k}} = \frac{\left[\begin{array}{c} r_{k-\ell_{1}} \\ \vdots \\ r_{k} \\ \vdots \\ r_{k} \end{array}\right]}{\mathbf{r}_{k}} = \frac{\left[\begin{array}{cccccc} f_{k-\ell_{1},0} & \dots & f_{k-\ell_{1},-\ell_{1}} & 0 & \dots & 0 \\ \vdots \\ f_{k-\ell_{1},\ell_{2}} & \dots & f_{k,0} & \dots & f_{k,-\ell_{1}} & 0 & \vdots \\ \vdots \\ \vdots \\ \vdots \\ r_{k} \\ \vdots \\ z_{k} \\ \vdots \\ z_{k+\ell_{1}+\ell_{2}} \end{array}\right]}{\mathbf{r}_{k}} = \frac{\left[\begin{array}{c} z_{k-\ell_{1}-\ell_{2}} \\ \vdots \\ z_{k} \\ \vdots \\ z_{k+\ell_{1}+\ell_{2}} \\ z_{k} \\ \vdots \\ z_{k+\ell_{1}+\ell_{2}} \\ z_{k} \\ z_{$$

or

_

_

$$\boldsymbol{r}_k = \boldsymbol{F}_k \boldsymbol{z}_k + \boldsymbol{n}_k, \quad \text{with} \quad \boldsymbol{n}_k \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{\Sigma} = \sigma_{\text{ch}}^2 \boldsymbol{I}).$$
 (5)

Let us assume that the modulation is BPSK. Hence, there is a simple mapping between the coded bits x_k and the modulated symbols z_k . To keep the notation simple, we will not use interleaved indices, and appropriate interleaving and deinterleaving should be assumed. Based on the extrinsic LLR of the coded bits provided by the channel decoder, $\{L_o^{m-1}(x_k)\}$, we first form soft estimates of the modulated symbols z_k as

$$\begin{split} \tilde{z}_{k} &= 1 P(z_{k} = 1) + (-1)P(z_{k} = -1) \\ &= 1P(x_{k} = 1) + (-1)P(x_{k} = 0) \\ &= 1 \frac{e^{L_{o}^{m-1}(x_{k})}}{1 + e^{L_{o}^{m-1}(x_{k})}} - 1 \frac{1}{1 + e^{L_{o}^{m-1}(x_{k})}} \\ &= \tanh\left(\frac{L_{o}^{m-1}(x_{k})}{2}\right). \end{split}$$
(6)

Define

$$\tilde{\boldsymbol{z}}_{k} \stackrel{\scriptscriptstyle\Delta}{=} [\tilde{z}_{t-\ell_{1}-\ell_{2}}, \dots, \tilde{z}_{t-1}, 0, \tilde{z}_{t+1}, \dots, \tilde{z}_{t+\ell_{1}+\ell_{2}}]^{T}$$
(7)

Then the soft estimate is used to cancel the intersymbol interference from the received signal r_k to obtain

$$\tilde{\boldsymbol{r}}_{k} \stackrel{\Delta}{=} \boldsymbol{r}_{k} - \boldsymbol{F}_{k} \tilde{\boldsymbol{z}}_{k} \tag{8}$$

$$= \boldsymbol{F}_k(\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_k) + \boldsymbol{n}_k \tag{9}$$

Next an instantaneous linear MMSE filter is applied to $\tilde{\boldsymbol{r}}_k$, to obtain

$$u_k = \boldsymbol{w}_k^H \tilde{\boldsymbol{r}}_k \tag{10}$$

where the filter $\boldsymbol{w}_k \in \mathbb{C}^{\ell_1+\ell_2+1}$ is chosen to minimize the expected value of mean-square error between the modulated symbol z_k and the filter output u_k :

$$\boldsymbol{w}_{k} = \arg\min_{\boldsymbol{w}\in\mathbb{C}^{\ell_{1}+\ell_{2}+1}} E\{|\boldsymbol{z}_{k} - \boldsymbol{w}^{H}\tilde{\boldsymbol{r}}_{k}|^{2}\}$$
$$= \arg\min_{\boldsymbol{w}\in\mathbb{C}^{\ell_{1}+\ell_{2}+1}} \boldsymbol{w}^{H}E\{\tilde{\boldsymbol{r}}_{k}\tilde{\boldsymbol{r}}_{k}^{H}\}\boldsymbol{w} - 2\Re[\boldsymbol{w}^{H}E\{\boldsymbol{z}_{k}\tilde{\boldsymbol{r}}_{k}\}] (11)$$

where E denotes expectation and \Re denotes the real part of a complex number. Note that

...

$$E\{\tilde{\boldsymbol{r}}_k \tilde{\boldsymbol{r}}_k^H\} = \boldsymbol{F}_k \boldsymbol{\Delta}_k \boldsymbol{F}_k^H + \boldsymbol{\Sigma}, \qquad (12)$$

$$E\{z_k \tilde{\boldsymbol{r}}_k\} = \boldsymbol{F}_k \boldsymbol{e},\tag{13}$$

where $\mathbf{\Delta}_k$ is defined as

$$\begin{aligned} \mathbf{\Delta}_{k} &\stackrel{\scriptscriptstyle\Delta}{=} \operatorname{cov}\{\mathbf{z}_{k} - \tilde{\mathbf{z}}_{k}\} = \operatorname{diag}\{1 - \tilde{z}_{t-\ell_{1}-\ell_{2}}^{2}, \dots, 1 - \tilde{z}_{t-1}^{2}, 1, \\ &\times 1 - \tilde{z}_{t+1}^{2}, \dots, 1 - \tilde{z}_{t+\ell_{1}+\ell_{2}}^{2}\} \end{aligned}$$
(14)

and e denotes a $[2(\ell_1 + \ell_2) + 1]$ -vector with all-zero entries, except for the $(\ell_1 + \ell_2 + 1)$ th entry, which is 1 (hence $F_k e$ is the $(\ell_1 + \ell_2 + 1)$ th column of F_k). The solution to (11) is given by

$$\boldsymbol{w}_k = (\boldsymbol{F}_k \boldsymbol{\Delta}_k \boldsymbol{F} + \boldsymbol{\Sigma})^{-1} \boldsymbol{F}_k \boldsymbol{e}$$
(15)

In order to form the LLR of the modulated symbol z_k , the instantaneous MMSE filter output u_k in (10), which represents a soft estimate of the bit z_k , is treated as a Gaussian random variable:

$$p(u_k \mid z_k) \sim \mathcal{N}_c(\mu_k z_k, \nu_k^2) \tag{16}$$

Conditioned on the modulated symbol z_k , the mean and variance of u_k are given by

$$\mu_{k} \stackrel{\Delta}{=} E\{u_{k} \mid z_{k}\}$$
$$= e^{T} F_{k}^{H} (F_{k} \Delta_{k} F_{k}^{H} + \Sigma)^{-1} F_{k} e \qquad (17)$$

$$\nu_k^2 \stackrel{\Delta}{=} \operatorname{var}\{u_k \mid z_k\} = E\{|u_k|^2 \mid z_k\} - \mu_k^2 = \boldsymbol{w}_k^H E\{\tilde{\boldsymbol{y}}_k \tilde{\boldsymbol{y}}_k^H\} \boldsymbol{w}_k - \mu_k^2$$
$$= \mu_k - \mu_k^2 \tag{18}$$

Note that the mean μ_k is real. Therefore the extrinsic information $L_i^m(x_k)$ delivered by the SISO equalizer is given by

$$L_{i}^{m}(x_{k}) = \log \frac{P(u_{k} \mid x_{k} = 1)}{P(u_{k} \mid x_{k} = 0)} = \log \frac{P(u_{k} \mid z_{k} = +1)}{P(u_{k} \mid z_{k} = -1)}$$
$$= -\frac{|u_{k} + \mu_{k}|^{2}}{v_{k}^{2}} + \frac{|u_{k} - \mu_{k}|^{2}}{v_{k}^{2}}$$
$$= \frac{4\Re\{\mu_{k}u_{k}\}}{v_{k}^{2}} = \frac{4\Re\{u_{k}\}}{1 - \mu_{k}}$$
(19)

For real-valued channels, $L_i^m(x_k) = \frac{2u_k}{1-\mu_k}$.

4.3. SISO Decoding

When convolutional codes are used as outer codes, the BCJR algorithm, any of its variants discussed in Section 4.1, or the SOVA can be used to generate soft output for the coded bits and the information bits. When parallel concatenated or serial concatenated convolutional codes (SCCC) are used, the Turbo decoding algorithm explained in Ref. 6 can be used. For low-density paritycheck codes, the belief propagation decoder [16] naturally produces soft output in order to iterate between the SISO equalizer and the decoder. When PCCC, SCCC, or LDPC outer codes are used, the decoding algorithm for softoutput decoding of the outer code itself is an iterative algorithm, which can be combined with the iterations between the decoder and the equalizer such as in Ref. 17.

We show some simulation results to demonstrate the efficiency of the Turbo equalization algorithm. Figure 4 shows the bit error rate as a function of the number of iterations for a 5-tap ISI channel with frequency response $F_1(z) = \sqrt{0.45} + \sqrt{0.25}z^{-1} + \sqrt{0.25}z^{-1}$ $\sqrt{0.15}z^{-2} + \sqrt{0.1}z^{-3} + \sqrt{0.05}z^{-4}$ when the outer code is a 16-state convolutional code with generator polynomials $[1 + D^2 + D^4, 1 + D + D^2 + D^4]$ and the block length is K = 2048 information bits. The equalizer and the decoder use the SOVA. The performance of the convolutional code in an AWGN channel is also shown. It can be seen that the bit error rate improves with iterations and the performance after 6 iterations is comparable to the performance of the code on an AWGN channel. Figure 5 shows the performance of a parallel concatenated outer code where the component codes are 16-state convolutional codes on the same channel for K = 5000. In every iteration of turbo equalization one iteration of Turbo decoding is performed within the outer decoder. The performance

can be seen to improve steadily with iterations and is within 0.8 dB from the capacity for this channel. These results show that turbo equalization is quite effective in removing ISI.

5. PERFORMANCE ANALYSIS

It is quite difficult to analytically predict the performance of the Turbo equalizer for a given interleaver. However, it is possible to derive bounds on the performance over the ensemble of all possible interleavers both when E_b/N_o is very high and when the length of the codewords $N \to \infty$. The following two types of analysis-distance spectrum based analysis and analysis of the iterative algorithm provide some insight into the performance of Turbo equalization.

5.1. Distance-Spectrum-Based Analysis

The main idea here is to treat the ISI channel as a convolutional code over complex (or real) field and, hence, to treat the overall system in Fig. 1 as a serial concatenated convolutional code (SCCC) where the error correction code is the outer code and the ISI channel is the inner code. Then, the performance of an ML decoder can be computed over the ensemble of all interleavers, according to the technique pioneered by Benedetto et al. [18]. Our explanation below is based on Ref. 18 but is adapted to the case when the inner code is an ISI channel.

Since data are transmitted in the form of blocks, we can think of the outer code and the ISI channel as equivalent block codes of codeword length N. For a given reference codeword (sequence), let $A^{C_o}(w, h)$ denote the total number of error sequences with input Hamming weight w and output Hamming weight h for the outer



Figure 4. Bit error rate performance with turbo equalization for 5-tap ISI channel and 16-state rate- $\frac{1}{2}$ convolutional outer code.



code. Each error sequence may result in one or more error events and h is the Hamming weight of the sum of all error events. When the outer code is a linear code, the all-zero sequence can be taken as the reference codeword. Then, the number of error sequences of a given w and h is the same as that of the number of codewords of the outer code with Hamming weight h and information weight w. Similarly, let $A^{C_{\text{ISI}}}(h, \delta^2)$ denote the average number of error sequences with input weight h and squared Euclidean distance (SED) δ^2 for the ISI channel. The average is over all possible reference sequences, since the SED corresponding to an error sequence depends on the reference codeword as well and, hence, the allzero sequence cannot be assumed as the reference. Techniques used to compute $A^{C_{\text{ISI}}}(h, \delta^2)$ can be found in the literature [19,20]. Under the assumption of a uniform interleaver, that is, over the ensemble of all interleavers, the average number of error sequences of input weight w and SED δ^2 , $A^{C_s}(w, \delta^2)$ can be shown to be [18]

$$A^{C_s}(w,\delta^2) = \sum_{h=d_f^o}^{N} \frac{A^{C_o}(w,h) \times A^{C_{ISI}}(h,\delta^2)}{\binom{N}{h}}$$
(20)

where N is the length of the interleaver and d_{f}^{o} is the free distance of the outer code. The probability of bit error under maximum-likelihood decoding can be upper bounded by using the union bound as

$$P_{b}(e) \leq \sum_{w=1}^{K} \frac{w}{K} \sum_{h} \sum_{\delta^{2} \in \Delta} A^{C_{s}}(w, \delta^{2}) Q\left(\sqrt{\frac{\delta^{2} \mathrm{RE}_{b}}{4N_{o}}}\right)$$
(21)

where Δ is the set of all possible squared Euclidean distances δ^2 and R is the rate of the outer code. This result should be interpreted with some caution. On one

Figure 5. Bit error rate performance with Turbo equalization for 5-tap ISI channel and 16-state rate- $\frac{1}{2}$ Turbo outer code.

hand, it is an upper bound computed over the ensemble of all possible interleavers and it is possible to design interleavers that perform better than the average. On the other hand, the Turbo equalization algorithm is not a true ML decoding algorithm and, hence, the performance can be worse than that predicted by (21). Nevertheless, the abovementioned approach can be used to derive some understanding of the effect of the interleaver and different parameters of the outer code. For large E_b/N_o , only the first few terms of the summation corresponding to small values of δ^2 and h dominate the performance. For small h, $A^{C_o}(w, h)$ and $A^{C_{ISI}}(h, \delta^2)$ can be approximated as [18]

$$A^{C_o}(w,h) \approx \sum_{n=1}^{n_{\max}^o(w)} T^{C_o}(w,h,n) \binom{N}{n}$$
(22)

$$A^{C_{ISI}}(h,\delta^2) \approx \sum_{n=1}^{n_{\max}^{\ell}(h)} T^{C_{ISI}}(h,\delta^2,n) \binom{N}{n}$$
(23)

where $T^{C_o}(w, h, n)$ is the number of error events for the outer code of input weight w and output weight h that are the concatenation of exactly n consecutive error events. By consecutive we mean that the error path diverges from the reference path and on merging back immediately diverges again until n such error events occur. Then, the paths remain merged until the end of the block. Similarly, let $T^{C_{\text{ISI}}}(h, \delta^2, n)$ be the number of error events of input weight *h* and output SED δ^2 that are the concatenation of n consecutive error events for the ISI channel. The quantities $n_{\max}^{o}(w)$ and $n_{\max}^{i}(h)$ refer to the maximum number of error events possible due an input error sequence of weight w and h for the outer code and ISI channel, respectively. It is important to note that for convolutional outer codes, the quantities $T^{C_o}(w, h, n)$ and $T^{C_{\text{ISI}}}(h, \delta^2, n)$ in (23) are independent of N. The probability of bit error can then be rewritten as

$$\begin{split} P_{b}(e) &\leq \sum_{w} \frac{w}{K} \sum_{h} \sum_{\delta^{2}} \sum_{n^{o}=1}^{n_{\max}^{o}(h)} \sum_{n^{i}=1}^{n_{\max}^{i}(h)} T^{C_{o}}(w,h,n^{o}) \\ &\times T^{C_{\mathrm{ISI}}}(h,\delta^{2},n^{i}) \times \frac{\binom{N}{n^{o}}\binom{N}{n^{i}}}{\binom{N}{h}} Q\left(\sqrt{\frac{\delta^{2}\mathrm{RE}_{b}}{4N_{o}}}\right) \end{split}$$
(24)

Using the approximation $\binom{N}{n} \approx \frac{N^n}{n!}$, and the fact that K = NR, the probability of bit error can be written as

$$\begin{split} P_{b}(e) &\leq \sum_{w} \frac{w}{R} \sum_{h} \sum_{\delta^{2}} \sum_{n^{o}=1}^{n_{\max}^{o}(h)} \sum_{n^{i}=1}^{n_{\max}^{i}(h)} T^{C_{o}}(w, h, n^{o}) \\ &\times T^{C_{\text{ISI}}}(h, \delta^{2}, n^{i}) N^{n^{o}+n^{i}-h-1} Q\left(\sqrt{\frac{\delta^{2} \text{RE}_{b}}{4N_{o}}}\right) \end{split}$$
(25)

where we can see that the probability of bit error depends on the interleaver length through the term $N^{n^o+n^i-h-1}$. Since the ISI channel is a non-recursive encoder, $n_{\max}^{i}(h) =$ h [6,18] and, hence, the exponent of N is always greater than or equal to zero. Therefore, the probability of bit error is at best independent of the length N of the codewords, which is the same situation as that for convolutional codes. This means that although, the system in Fig. 1 is a serial concatenated convolutional code, there is no interleaving gain when the outer code is a simple convolutional code, since the ISI channel is a nonrecursive inner code. The interleaver is still useful since the h different error events each correspond to an SED of at least δ^2_{\min} and, hence, the overall free distance can be up to $d_f^o \times \delta_{\min}^2$. The interleaver is also required to break up the correlation at the output of the equalizer.

This tells us that in order to achieve close to capacity performance, the outer code should be a Turbo code or an LDPC code, in which case, the performance of the code gets better with increasing N and, hence, the overall performance improves with increasing N.

5.1.1. Binary Precoding. It is known from Benedetto et al. [18] that an interleaving gain is possible even with a simple convolutional outer code if the inner encoder is recursive. It is possible to make the ISI channel appear recursive to the outer code by encoding the interleaved output y in Fig. 1 by a rate-1 recursive convolutional encoder as shown in Fig. 6. The precoder is a recursive encoder with generator polynomial 1/g(D) with $g(D) = \frac{1}{2}$

 $\sum_i g_i D^i$, where J is the number of memory elements in the

precoder and $g_i \in \{0, 1\}$. The outputs of the precoder are transmitted over the ISI channel after modulation. This type of precoding, which we refer to as *binary precoding*, should not be confused with the conventional type of precoding such as Tomlinson-Harashima (TH) precoding, which is intended to cancel ISI at the transmitter. Unlike TH precoding, in binary precoding, channel knowledge is not required at the transmitter and the purpose is only to make the channel trellis appear recursive. When the modulation is memoryless, the trellis of the precoder and that of the ISI channel can be combined into one trellis as shown in Fig. 7. If $J \leq L$, then the resulting combined trellis has the same number of states as that of the ISI channel and, hence, there is no increase in the equalization complexity. However, since the precoder is recursive, the combination of the channel and the precoder becomes recursive and, hence, a significant interleaving gain is possible. From the combined transversal filter, the trellis diagram can be easily found. An example of a precoded 2-tap ISI channel with g(D) = 1 + D and the associated trellis diagram is shown in Fig. 8. The ISI channel considered is the same as that in Fig. 3 with transfer function $F_2(z) = f_0 + f_1 z^{-1}$.



Figure 6. System model with binary precoding.



Figure 7. Transversal filter for precoded ISI channel.

Note from the trellis in Fig. 8 that for the precoded channel, a weight-1 error sequence produces an infinite-length error event and that a minimum input weight of 2 is required to produce a finite-length error event. Therefore, an input error sequence with weight h can produce a maximum of $\lfloor h/2 \rfloor$ error events and $n_{\max}^i(h) = \lfloor h/2 \rfloor$ in (25). Therefore, the maximum exponent of N is $\lfloor h/2 \rfloor - h$ and since $h \ge d_f^o$, the exponent of N is always negative if $d_f^o \ge 3$. Thus the BER in (25) decreases with increasing interleaver length; this phenomenon has been termed as interleaving gain [18].

Since the free distance of the outer code needs to be greater than or equal to only 3, even simple codes such as convolutional codes with small constraint length usually perform very well with binary precoding. Thus significant reduction in complexity can be obtained compared to using Turbo outer codes such as in Ref. 21 or, better performance can be obtained compared to using convolutional outer codes without precoding such as in Ref. 3. Figure 9 shows the bit error rate performance



Figure 8. (a) Two-tap ISI channel with precoding; (b) trellis of precoded channel.



Figure 9. Bit error rate performance for a 5-tap static ISI channel with and without binary precoding.

on a 5-tap ISI channel with transfer function $F_1(z) = \sqrt{0.45} + \sqrt{0.25z^{-1}} + \sqrt{0.15z^{-2}} + \sqrt{0.1z^{-3}} + \sqrt{0.05z^{-4}}$ for a block length of K = 5000 bits. The outer code used is a simple rate- $\frac{1}{2}$, 2-state convolutional code with generator polynomials [1, 1 + D]. The precoder used is 1/(1 + D) and the error performance is shown for different iterations up to 18 iterations. Also shown for comparison is the performance of a 16-state Turbo code (parallel concatenated code) with 12 iterations. It can be seen that even a simple 2-state outer code can provide performance within 0.3 dB as that of a 16-state Turbo code; but the decoding complexity for the former is significantly lesser compared to that of the Turbo code.

The combination of the outer code, interleaver, and the rate-1 recursive precoder can be thought of as a serial concatenated outer code with a recursive inner encoder whose output is transmitted over the ISI channel and that the interleaving gain is really due to the concatenated outer code. However, it should be noted that there is no interleaver between the precoder and the ISI channel and, hence, a separate decoder is not required for the precoder. By combining the trellis of the precoder and the channel, the equalization complexity is reduced and, hence, it is useful to think of this technique as a precoding technique rather than a mere concatenated outer code. Analysis of precoded ISI channels based on the distance spectrum for partial response channels can be found elsewhere [22-24]. The design of precoders based on optimizing the distance spectrum has been considered by Lee [25].

5.2. Analysis of the Iterative Equalization Algorithm

The aforementioned analysis is based on the assumption of an ML decoder whereas the Turbo equalization algorithm is not an ML decoding algorithm and, therefore, the performance of the Turbo equalization can be quite different from that predicted by the analysis above. In order to accurately characterize the iterative process, we need to determine how the extrinsic information evolves from one iteration to another. Since the extrinsic information vector is an N-dimensional vector, it is quite difficult to characterize the evolution of this vector. When the length $N \to \infty$, we can assume that the extrinsic information $(\ldots, L_i^m(x_k), \ldots)$ or $(\ldots, L_o^m(x_k), \ldots)$ are sequences of independent identically distributed random variables and, hence, only the PDFs of $L_i^m(x_k)$ and $L_{\alpha}^{m}(x_{k})$ need to be tracked from one iteration to another. It is still quite difficult to compute the pdf as a function of iterations. Ten Brink [26] and El Gamal and Hammons [27] showed that for Turbo codes, if the PDF is assumed to be Gaussian and only a single parameter related to the PDF is tracked, the approximation is still quite good. This idea was extended to analyze Turbo equalization in Ref. 28. Here, we will explain how to use this technique to analyze the performance of Turbo equalization.

We first introduce the concept of equivalent channels. Consider the interleaved coded sequence $\mathbf{y} = \{y_k\}$ and BPSK modulation with $z_k = (2y_k - 1)$. During the *m*th iteration in the Turbo equalization algorithm, the inner equalizer provides LLRs (or extrinsic information) $L_i^m(x_k)$. This LLR can be thought of as the output of a hypothetical channel with binary inputs as "seen" by the outer decoder. Let us define two quantities for this equivalent channel:

$$\mu_i^m \stackrel{\Delta}{=} E_{y_k}[(2y_k - 1)L_i^m(y_k)] \tag{26}$$

$$(\sigma_i^m)^2 \stackrel{\Delta}{=} E_{y_k}[((2y_k - 1)L_i^m(y_k))^2] = E_{y_k}[(L_i^m(y_k))^2] \quad (27)$$

A measure of reliability of this equivalent channel is the ratio $\text{SNR}_i^m = \left(\frac{\mu_i^m}{\sigma_i^m}\right)^2$, which can be thought of as the

SNR of the channel as seen by the outer decoder during the *m*th iteration. The higher SNR_i^m is, the more reliable the effective channel as seen by the outer decoder during the *m*th iteration is. Similarly, the outer decoder produces extrinsic information $L^m_{\alpha}(y_k)$ during the *m*th iteration and a similar SNR, SNR_o^m can be defined as the SNR of the equivalent channel as seen by the equalizer during the mth iteration. During the mth iteration, the inner SISO equalizer is provided with an equivalent channel with SNR SNR_{o}^{m-1} by the outer code and an AWGN channel with variance σ_{ch}^{2} . At the output, the equalizer produces an equivalent channel with $SNR SNR_i^m$. The outer decoder in turn observes this channel and increases the equivalent SNR at the output of the decoder to SNR_{0}^{m} . Let us denote the SNR at the output of the equalizer by the transfer function $\mathcal{F}_i(p,q)$, where p is the input SNR from the outer decoder and $q = \sigma_{ch}^2$ is the variance of the AWGN. Similarly, let $\mathcal{F}_o(p)$ denote the equivalent SNR at the decoder output when the input SNR is *p*. Then, we have

$$SNR_i^m = \mathcal{F}_i(SNR_o^{(m-1)}, \sigma_{ch}^2)$$
(28)

$$\mathrm{SNR}_o^m = \mathcal{F}_o(\mathrm{SNR}_i^m) \tag{29}$$

The evolution of the equivalent SNR with iterations is best illustrated with the help of a diagram such as the one suggested by Ten Brink [26] and shown in Figs. 10 and 11. Figure 10 shows two curves: the function $\mathcal{F}_o(p)$ and the function $\mathcal{F}_i^{-1}(p, \sigma_{ch}^2)$ for a fixed σ_{ch}^2 [so we refer to this simply as $\mathcal{F}_i^{-1}(p)$]. The function $\mathcal{F}_i^{-1}(p)$ is represented by simply drawing the function $\mathcal{F}_i(p)$ with the x and y axes inverted. Since the output of one of the SISOs is the input to the other, by drawing the curves with



Figure 10. Graphical demonstration of convergence of iterative equalization and decoding: serial concatenation between a 2-state convolutional outer code and a 5-tap ISI channel.


Figure 11. Graphical demonstration of convergence of iterative equalization and decoding: serial concatenation between a 2-state convolutional outer code and a precoded 5-tap ISI channel.

the axis reversed for one of the curves, the evolution of the SNRs can be traced by drawing vertical and horizontal lines between the two curves. That is, the iterations begin at the point $\text{SNR}_i^{(1)}$ and a vertical line from the curve $\mathcal{F}_i^{-1}(p)$ to the curve $\mathcal{F}_o(p)$ followed by a horizontal line to the curve $\mathcal{F}_i^{-1}(p)$ denotes the change of the equivalent SNR during consistencies. the equivalent SNR during one iteration. The outer code in this case is a 2-state convolutional code and the channel is the 5-tap ISI channel with frequency response $F_1(z) =$ $\sqrt{0.45} + \sqrt{0.25}z^{-1} + \sqrt{0.15}z^{-2} + \sqrt{0.1}z^{-3} + \sqrt{0.05}z^{-4}$ and the $E_b/N_o = 3.6$ dB. It can be seen from Fig. 10 that the two curves intersect, which represents a fixed point for the iterations and the equivalent SNR cannot improve beyond the fixed point. The achievable bit error rate can be computed from the equivalent SNR corresponding to the fixed point. If the two curves $\mathcal{F}_o(p)$ and $\mathcal{F}_i^{-1}(p)$ do not intersect, then the equivalent SNR steadily increases to infinity, which denotes correct decoding or convergence of the turbo equalization algorithm to the correct solution. Such a situation occurs for the same convolutional code and same E_b/N_o , when precoding is used with the channel and is shown in Fig. 11.

Note that in order for the iterative procedure to converge to the correct decision, it is not necessary that both SNR_i^m and SNR_o^m go to infinity; it is sufficient that any one of them, SNR_i^m or SNR_o^m tends to ∞ , since the SNRs are defined using the extrinsic LLRs only. The shape of $\mathcal{F}_i^{-1}(p)$ depends on the σ_{ch}^2 (and, hence, on E_b/N_o). The minimum E_b/N_o for which $\mathcal{F}_i^{-1}(p, E_b/N_o)$ does not intersect $\mathcal{F}_o(p)$ is called the *threshold*. In general, for finite impulse response ISI channels, $\mathcal{F}_i^m(\infty, \sigma_{\text{ch}}^2)$ is finite and, hence, in order to get arbitrarily small probability of error for a fixed σ_{ch}^2

capacity approaching codes such as Turbo codes or LDPC codes must be used. With convolutional outer codes, the bit error rate cannot be made arbitrarily small. However, with precoding $\mathcal{F}_i^m(\infty, \sigma_{\mathrm{ch}}^2) \to \infty$ and, hence, even simple convolutional codes can be used to obtain arbitrarily small probability of error.

Finally, it should be noted that parameters other than the equivalent SNR have been used to characterize the iterative process. Ten Brink used mutual information [26], and Narayanan has used the correlation between the equivalent soft output and the actual coded bit [28].

5.2.1. Computing the Transfer Functions. For trellisbased equalizers such as the BCJR algorithm or SOVA, it is quite difficult to analytically compute the functions \mathcal{F}_i ; hence, Monte Carlo simulations have to be used and ensemble averages in (27) are replaced by a time average. In order to simulate the input extrinsic information, $L_o^m(y_k)$ can be assumed to be a Gaussian random variable whose variance is twice the mean. This assumption has been shown to be reasonably accurate [29]. The following procedure is then used to evaluate $\mathcal{F}_i(p,q)$:

- 1. Draw a sequence of random i.i.d. bits $y_k \in \{0, 1\}$, set $z_k = 2y_k 1$ for k = 1, 2, ..., N
- 2. Compute $r_k = \sum_{i=-l_1}^{l_2} f_{k-i} z_k + n_k$ where n_k are samples of i.i.d. Gaussian random variables with zero mean and variance q.
- 3. Draw a sequence of random variables for $L_o^{(m-1)}(y_k)$ from the distribution $\mathcal{N}(2pz_k, 4p)$, for k = 1, 2, ..., N.
- 4. Compute the equalizer output $L_i^m(y_k)$, k = 1, 2, ..., N.

5. Compute SNR_i^m =
$$\frac{\left(\frac{1}{N}\sum_{k=1}^{N} z_k L_i^m(y_k)\right)^2}{\frac{1}{N}\sum_{k=1}^{N} (z_k L_i^m(y_k))^2}$$

When MMSE equalizers are used such as the one described in Section 4.2, it is possible to obtain the output SNR_i^m semianalytically. That is, we do not have to simulate the equalizer. The following procedure can be used to compute $\mathcal{F}_i(a, b)$ —given the channel $\mathbf{F}_t = \mathbf{F}$, the noise variance $\sigma_{\text{ch}}^2 = q$:

1. For
$$i = 1, 2, ..., N$$
 draw i.i.d. $L_i^m(y_k) \sim \mathcal{N}(2az_k, 4a)$
2. Let $\mathbf{\Delta}_k \stackrel{\Delta}{=} \operatorname{diag} \left\{ 1 - \operatorname{tanh} \left(\frac{L_i(y_{k-\ell_1-\ell_2})}{2} \right)^2, ..., 1 - \operatorname{tanh} \left(\frac{L_i(y_{k-1})}{2} \right)^2, 1, 1 - \operatorname{tanh} \left(\frac{L_i(y_{k+1})}{2} \right)^2, ..., 1 - \operatorname{tanh} \left(\frac{L_i(y_{k+\ell_1+\ell_2})}{2} \right)^2 \right\}.$

3. Compute

$$\mu_{k} \stackrel{\Delta}{=} \boldsymbol{e}^{T} \boldsymbol{F}^{H} \left(\boldsymbol{F} \boldsymbol{\Delta}_{k} \boldsymbol{F}^{H} + \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{F} \boldsymbol{e}$$
(30)

$$\operatorname{SNR}_{i}^{m} \stackrel{\scriptscriptstyle \Delta}{=} \frac{1}{N} \sum_{k=1}^{N} \frac{1}{2} \frac{2q\mu_{k}}{1-\mu_{k}}$$
(31)

where q = 1 for real channels and q = 2 for complex channels. The transfer function of the outer code $\mathcal{F}_o(a)$ cannot be computed analytically and has to be computed via Monte Carlo simulations using the following procedure:

- 1. Draw a sequence of random i.i.d. bits $y_k \in \{0, 1\}$.
- 2. Draw a sequence of random variables for $L_i^m(y_k)$ from the distribution $\mathcal{N}(2az_k, 4a)$, for k = 1, 2, ..., N.
- 3. Compute the decoder output $L_o^m(\mathbf{y}_k), \ k = 1, 2, \dots, N$.

4. Compute SNR_o^m =
$$\frac{\left(\frac{1}{N}\sum_{k=1}^{N} z_k L_o^m(y_k)\right)^2}{\frac{1}{N}\sum_{k=1}^{N} (z_k L_o^m(y_k))^2}.$$

6. MORE RECENT RESULTS

Turbo equalization is an area of active research and new results on theory and application are emerging. Turbo equalization with carefully designed LDPC outer codes has been shown to provide close to capacity performance. The design of LDPC codes with different kinds of equalizers is an area of current research. Reduced-complexity Turbo equalization, adaptive Turbo equalization, and applications to multiantenna systems and other wireless systems [30,31] are few of the areas under study.

Acknowledgments

The author would like to thank Dũng Ngoc Doan for help with Figs. 10 and 11 and for proofreading the manuscript. The author would also like to thank Prof. Xiaodong Wang for help with the material in section 4.2.

BIOGRAPHY

Krishna R. Narayanan received his Ph.D. degree from Georgia Institute of Technology, Atlanta, in 1988. Since then he has been an assistant professor in the Department of Electrical Engineering at Texas A&M University. His research interests are mainly in the areas of advanced modulation, coding and receiver design for wireless communications, and magnetic recording. Specifically, he has worked in the areas of turbo coding, and iterative signal processing. He currently serves as an associate editor for *IEEE Communication Letters* and an editor for *IEEE Transactions on Wireless Communications*. He is a recipient of the CAREER award from the National Science Foundation in 2001.

BIBLIOGRAPHY

- 1. J. G. Proakis, Digital Communications, McGraw-Hill, 2001.
- A. Duel-Hallen and C. Heegard, Delayed decision feedback sequence estimation, *IEEE Trans. Commun.* 37: 428-436 (May 1989).
- C. Douillard, Iterative correction of intersymbol interference: Turbo-equalisation, *European Trans. Commun.* 507–511 (Sept.–Oct. 1995).
- C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding, *Proc. IEEE Int. Conf. Communication*, (Geneva, Switzerland), June 1993, pp. 1064–1070.
- L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* 20: 284–287 (March 1974).
- 6. J. Proakis, ed., *Wiley Encyclopedia in Telecommunications*, "Concatenated coding and iterative decoding," Wiley, 2002.
- P. Roberston, E. Villebrun, and P. Hoeher, A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain, *Proc. IEEE Int. Conf. Communication*, 1995, pp. 1009-1013.
- 8. J. Hagenauer and P. Hoher, A Viterbi algorithm with soft-decision outputs and its applications, *Proc. IEEE GLOBECOM*, (Dallas, TX), Nov. 1989, pp. 47.1.1-47.1.7.
- 9. L. Pake and P. Robertson, Improved decoding with SOVA in a parallel concatenated (turbo-code) scheme, *Proc. IEEE Int. Conf. Communication*, 1996, pp. 102–106.
- V. Franz and J. B. Anderson, Concatenated decoding with a reduced search BCJR algorithms, *IEEE J. Select. Areas Commun.* 16: 186–195 (Feb. 1998).
- K. R. Narayanan, U. D. Gupta, and B. Lu, Low-complexity iterative equalization and decoding with binary precoding, *Proc. IEEE Int. Conf. Communication*, 2000, pp. 1–5.
- 12. S. H. Muller, W. H. Gerstacker, and J. B. Huber, Reduced state soft output trellis equalization incorporating soft

feedback, Proc. IEEE Global Communication Conf., 1996, pp. 95–100.

- A. Glavieux, C. Laot, and J. Labat, Turbo equalization over a frequency selective channel, *Proc. Int. Symp. Turbo Codes* and *Related Topics*, Sept. 1997, pp. 96–102.
- X. Wang and H. Poor, Iterative (turbo) soft interference cancellation and decoding for coded CDMA, *IEEE Trans. Commun.* 47: 1046-1061 (July 1999).
- M. Tüchler, R. Koetter, and A. Singer, Iterative correction of isi via equalization and decoding with priors, *Proc. IEEE Int. Symp. Information Theory*, 2000, p. 100.
- D. MacKay and R. M. Neal, Near shannon limit performance of low density parity check codes, *Electron. Lett.* 457–458 (March 1996).
- J. Fan, E. Kurtas, A. Friedmann, and S. W. McLaughlin, Low density parity check codes for digital magnetic recording, *Proc. Allerton Conf. Communication Control and Computing*, 1999.
- S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Design and performance analysis, *IEEE Trans. Inform. Theory* 42: 409–429 (April 1998).
- E. Zehavi and J. K. Wolf, On the performance evaluation of trellis codes, *IEEE Trans. Inform. Theory* 33: 483–501 (July 1987).
- S. A. Raghavan, J. K. Wolf, and L. B. Milstein, On the performance evaluation of isi channels, *IEEE Trans. Inform. Theory* 39: 957-965 (1993).
- 21. D. Raphaeli and Y. Zarai, Combined turbo equalization and turbo decoding, *IEEE Commun. Lett.* 2: 107-109 (April 1998).
- T. Souvignier et al., Turbo decoding for PR4: Parallel versus serial concatenation, *Proc. IEEE Int. Conf. Communication*, 1999, pp. 1638-1642.
- T. Duman and E. Kurtas, Performance bounds for high rate linear codes over partial response channels, *IEEE Trans. Inform. Theory* 47: 1201–1205 (March 2001).
- 24. L. McPheters, S. W. McLaughlin, and K. R. Narayanan, Precoded PRML, serial concatenation, and iterative (turbo) decoding for digital magnetic recording, *IEEE Trans. Magn.* 2325-2327 (Sept. 1999).
- I. Lee, The effect of a precoder on serially concatenated coding system with ISI channel, *IEEE Trans. Commun.* 1168–1175 (July 2001).
- S. Tenbrink, Iterative decoding for multicode CDMA, Proc. IEEE Vehicular Technology Conf., 1999, pp. 1876–1880.
- 27. H. El Gamal and R. Hammons, Analyzing the turbo decoder using the gaussian assumption, *Proc. Conf. Information Sciences and Systems*, March 2000.
- K. R. Narayanan, Effect of precoding on the convergence of turbo equalization for partial response channels, *IEEE J. Select. Areas Commun.* 19: 686–698 (April 2001).
- 29. D. Divsalar, S. Dolinar, and F. Pollara, Iterative turbo decoder analysis based on density evolution, *IEEE J. Select.* Areas Commun. 891–907 (Feb. 2001).
- 30. G. Bauch, H. Khorram, and J. Hagenauer, Iterative equalization and decoding in mobile communications systems, *Proc. Eur. Personal Mobile Communication Conf*, 1997, pp. 307–312.

 G. Bauch and V. Franz, Iterative equalization and decoding for the GSM system, *Proc. IEEE Vehicular Technology Conf.* (Ottawa), May 1998, pp. 2262–2266.

TURBO PRODUCT CODES FOR OPTICAL CDMA SYSTEMS

STEVEN W. McLAUGHLIN CENK ARGON Georgia Institute of Technology Atlanta, Georgia

1. INTRODUCTION

There exist various access schemes for participants of an optical fiber communications system [1]:

- 1. Time-division multiple access (TDMA)
- 2. Wavelength-division multiple access (WDMA)
- 3. Code-division multiple access (CDMA)

In TDMA [2] systems, each user (or transmitter) is assigned a specific time slot for data transmission. On the other hand, a specific wavelength is reserved for each transmitter in WDMA [1] systems. CDMA [2] is a different approach from TDMA or WDMA. In CDMA, not a wavelength or a time slot, but a specific pseudorandom address sequence is assigned to each participant. This is advantageous if compared to TDMA and WDMA because pseudorandom address sequences enable data transmission of users at overlapping times and wavelengths. Other advantages of CDMA include [2] (1) security against unauthorized users, (2) protection against jamming, (3) flexibility of adding users, and (4) asynchronous access capability.

In an optical CDMA system [3,4], an optical orthogonal code (OOC) sequence is assigned to each user and data are sent via the destination user's OOC sequence. Different from electrical pseudorandom sequences, which can consist of bipolar pulses, OOC sequences are unipolar, where instead of the -1 and +1 levels, only the 0 and +1 levels are available in intensity-modulated, direct-detection optical systems. Because of this, true orthogonality cannot be achieved. Therefore, OOC sequences have to be very long to support large numbers of users, and this introduces large bandwidth expansion [2].

The employment of error correction codes (ECCs) is possible in order to improve the performance (and hence to reduce the bandwidth expansion factor) of optical CDMA systems [5–7]. Zhang [5] proposes the use of asymmetric ECCs for optical CDMA with ON/OFF keying (OOK), whereas Kim and Poor [6] and Ohtsuki and Kahn [7] suggest using parallel concatenated convolutional codes (PCCC), namely, Turbo codes [8,9] in optical CDMA with pulse position modulation (PPM). This article presents the application of Turbo product codes (TPC) [17] for performance improvement of optical CDMA systems with either OOK or binary PPM.

2728 TURBO PRODUCT CODES FOR OPTICAL CDMA SYSTEMS

TPCs have near-optimum performance [17] like Turbo codes. Furthermore, in terms of storage requirements and number of operations, TPCs have considerably less complexity if compared to Turbo codes. This makes TPCs very attractive for various applications that require high performance at high code and data rates. For example, the application of TPCs is considered in Ref. 26 for optical systems employing dense wavelength-division multiplexing (DWDM).

The organization of this article is as follows: Section 2 summarizes the TPC decoding algorithm for binary memoryless channels. Section 3 describes optical CDMA systems and their channel models. The application of TPCs in optical CDMA and performance curves obtained via simulations are given is Section 4 and Section 5, respectively. Finally, we conclude with some remarks in Section 6.

2. TURBO PRODUCT CODES FOR BINARY MEMORYLESS CHANNELS

In Sections 2.1–2.4 we give some background on product codes and in Section 4 describe their use in optical CDMA systems.

2.1. Product Code Construction

Product codes, or "iterated" codes, are in the class of linear block codes [21], which are widely used for forward error correction (FEC) in many of today's communications systems. Product codes were first introduced by Elias [18] in 1954, and their construction is basically an efficient way for building long block codes from two or more shorter block codes. The construction of a two-dimensional product code can be described as follows: Consider a $k_1 \times k_2$ array of information bits as shown in Fig. 1. First, the columns are encoded using a systematic (n_1, k_1, d_1) linear code C_1 ; then, the rows are encoded using a systematic (n_2, k_2, d_2) linear code C_2 . Here, n_i, k_i , and $d_i (i = 1, 2)$ are the codeword length, the number of information bits, and the minimum Hamming distance, respectively. The linear codes C_1 and C_2 are called constituent or component codes and the resultant (n_1n_2, k_1k_2, d_1d_2) product code has a code rate of $R_c = (k_1 k_2)/(n_1 n_2)$. The idea of a product code



Figure 1. Product code construction.

can be further expanded to dimensions greater than 2. Nevertheless, to reduce encoding and decoding complexity, two-dimensional product codes are usually considered.

Product codes can be decoded by a hard-decision row decoder followed by a hard-decision column decoder (or vice versa). On reception of a noisy data vector, a hard-decision decoder strictly decides on bit 1 or 0 for each component of this vector and operates on these values. This is a lowcomplexity decoding method; nevertheless, for better performance, an iterative soft-input/soft-output (SISO) (i.e., soft-decision) decoding algorithm has to be used [17,19]. Different from the hard-decision decoder (which uses a threshold device), a soft-decision decoder uses the raw received analog signal to obtain reliability information on each bit position. The block diagram of an iterative SISO decoder for a product code is shown in Fig. 2.

On reception of the noisy data matrix R, the SISO row decoder calculates extrinsic (reliability) information matrix $W^{(1)}$ and passes this as a priori information to the SISO column decoder. The extrinsic information about a bit position is obtained via the help of all other bit positions, as described in the next section. After obtaining the information from the SISO row decoder, the SISO column decoder in turn calculates extrinsic information matrix $W^{(2)}$ and passes this back as a priori information to the row SISO decoder. Decoding continues in an iterative fashion until either estimate $R^{(1)}$ or $R^{(2)}$ is assigned as the decoded matrix. One efficient method of iterative SISO decoding is the Turbo product decoding algorithm proposed by Pyndiah [17] that is described and adapted for the binary memoryless channel model in the following section.

2.2. Extraction of Soft Information from a Binary Memoryless Channel

As we will observe later, optical CDMA systems can be modeled as binary memoryless channels [2]. Hence, there is a need to adapt a method for extracting soft information for these channels. That is, even though the channel is a "hard" channel that uses threshold detection, we can extract certain "soft" information given the error probability of the channel. It will take some time to explain this next. Let us assume that $X = x_0 x_1 \cdots x_{n-1}$ denotes the transmitted codeword and $R = r_0 r_1 \cdots r_{n-1}$ denotes the received vector, namely, a possibly noise corrupted version of X. The reliability for the *j*th component of R is given by the loglikelihood ratio (LLR):

$$\Lambda(r_j) = \log\left[\frac{\Pr(x_j = 1 \mid r_j)}{\Pr(x_j = 0 \mid r_j)}\right]$$
(1)

For $Pr(x_j = 1) = Pr(x_j = 0) = \frac{1}{2}$, Eq. (1) can be expressed using the transition probabilities of the binary memoryless channel and is found to be

$$\Lambda(r_j) = \log\left[\frac{\Pr(r_j \mid x_j = 1)}{\Pr(r_j \mid x_j = 0)}\right].$$
(2)

The Turbo product decoding algorithm applies the Chase algorithm [20], a suboptimum maximum-likelihood decoding method for a linear (n, k, d) block code C, iteratively to the columns and rows of the product



Figure 2. Iterative soft-input/soft-output decoder.

codeword. The Chase algorithm can be described as follows. First, the f least reliable bit positions of R are determined and 2^f test sequences are formed by perturbing these positions. Here, perturbation means using all possible combinations of zeros and ones in the least reliable bit positions. The parameter f is chosen as $f \ll k$. After perturbation of the least reliable bit positions, the 2^f test sequences are passed through an algebraic decoder for linear code C. The codewords obtained via the algebraic decoder are called *candidate codewords* and are denoted by $C^i(i = 1, 2, ..., 2^f)$. The distance between candidate codeword $C^i = c_0^i \ c_1^i \ \cdots \ c_{n-1}^i (c_j^i \in \{0, 1\})$ and received vector R is found by using the following metric, called "analog weight" [20]:

$$l(R, C^{i}) = \sum_{j=0}^{n-1} \left| (r_{j} \oplus c_{j}^{i}) \Lambda(r_{j}) \right|$$
(3)

where \oplus denotes modulo-2 addition. This metric is calculated for all candidate codewords. The final output of the Chase decoder is the decoded codeword, which is the candidate codeword closest to received vector R. If C^{D} denotes the decoded codeword, then the following condition is used to determine C^{D} :

$$l(R, C^{D}) \le l(R, C^{i}) \text{ for all } i$$
(4)

2.3. Computation of Extrinsic Information

Once a decision C^D is made, soft output (or extrinsic information) for each bit position c_j^D is needed to be passed to the next decoding stage. The reliability information of c_j^D is given by the LLR of the transmitted symbol x_j and can be written as

$$\Lambda(c_j^D) = \log\left[\frac{\Pr(x_j = 1 \mid R)}{\Pr(x_j = 0 \mid R)}\right]$$
(5)

Here, the computation of the LLR is different than the LLR in Eq. (1), since it takes into account that C^{D} is one of the 2^{k} codewords of linear code *C*. The probabilities in Eq. (5) can be expressed as

$$\Pr(x_j = \nu \mid R) = \sum_{C^i \in S_j^{\nu}} \Pr(X = C^i \mid R)$$
(6)

where S_j^{ν} denotes the set of all candidate codewords with $\nu \in \{0, 1\}$ at their *j*th bit position. Applying Bayes' rule [21]

and assuming that $Pr(X = C^i)$ to be equal for all *i*, the LLR in (5) can be written as

$$\Lambda(c_j^D) = \log \left[\frac{\sum_{\substack{C^i \in S_j^1}} \Pr(R \mid X = C^i)}{\sum_{\substack{C^i \in S_j^0}} \Pr(R \mid X = C^i)} \right]$$
(7)

Assuming independent and identically distributed (i.i.d.) signal components r_i , $\Pr(R \mid X = C^i)$ can be expressed as

$$\Pr(R \mid X = C^{i}) = \prod_{m} \Pr(r_{m} \mid x_{m} = c_{m}^{i})$$
(8)

Using Eq. (8) in Eq. (7), we obtain the following expression for the LLR:

$$\Lambda(c_j^D) = \Lambda(r_j) + w_j \tag{9}$$

where w_i is defined as the extrinsic information given by

$$w_{j} = \log \left[\frac{\sum\limits_{C^{i} \in S_{j}^{1}} \prod\limits_{m \neq j} \Pr(r_{m} \mid x_{m} = c_{m}^{i})}{\sum\limits_{C^{i} \in S_{j}^{0}} \prod\limits_{m \neq j} \Pr(r_{m} \mid x_{m} = c_{m}^{i})} \right]$$
(10)

As can be observed, the extrinsic information for bit position *j* is calculated via information obtained from all bit positions except bit position *j* itself.

The extrinsic information w_j , calculated as in Eq. (10), requires all candidate codewords. To reduce complexity, the following approach is used. Let $C^{\min(1),j} \in S_j^1$ be the codeword closest to R with a 1 in its *j*th bit position. Similarly, $C^{\min(0),j} \in S_j^0$ is the closest codeword to R with a 0 in its *j*th bit position. Using these closest (and hence dominant) codewords, the extrinsic information w_j can be approximated by

$$w_j \approx \log \left[\frac{\prod_{m \neq j} \Pr(r_m \mid x_m = c_m^{\min(1), j})}{\prod_{m \neq j} \Pr(r_m \mid x_m = c_m^{\min(0), j})} \right]$$
(11)

The calculation in this equation is not as accurate as the expression in Eq. (10). However, it has reduced complexity, since only two candidate codewords need to be considered to obtain the extrinsic information.

2730 TURBO PRODUCT CODES FOR OPTICAL CDMA SYSTEMS

2.4. Turbo Decoding of Product Codes

Having defined the computation of the extrinsic information, we are now able to summarize the Turbo decoding algorithm for a binary memoryless channel:

- 1. The Chase algorithm is applied to the rows of the product codeword and decision $C^D = c_0^D c_1^D \cdots c_{n-1}^D$ is obtained for each row.
- 2. $C^{\min(1),j}$ and $C^{\min(0),j}$ are determined among the candidate codewords. In fact, one of these was already found in the previous step; thus, C^D is equal to either $C^{\min(1),j}$ or $C^{\min(0),j}$.
- 3. If both $C^{\min(1),j}$ and $C^{\min(0),j}$ exist among the candidate codewords, then the extrinsic information w_j is evaluated using Eq. (11). If one of these codewords cannot be found, then the extrinsic information is approximated as

$$w_j \approx \beta (2c_j^D - 1) \tag{12}$$

where $\beta \ge 0$ is a predetermined reliability factor that increases with each iteration.

- 4. The extrinsic information w_j is scaled so that the average of $|w_j|$ is equal to one. The reason for this will be explained later.
- 5. The reliability information of r_j is updated as

$$\Lambda(r_j) = \Lambda_j + \alpha w_j \tag{13}$$

where Λ_j is the reliability of the original received *j*th symbol and $\alpha \ge 0$ is a weight factor to combat high standard deviation in w_j and high BER during the first iterations.

6. The product codeword with the updated reliability information is the input to the next decoding stage. The above-mentioned procedure is applied this time to the columns of the product codeword. Decoding continues in an iterative fashion (i.e., row decoding → column decoding → row decoding → ...) until user-defined termination.

One iteration of the preceding algorithm means row decoding (a half-iteration) followed by column decoding (another half-iteration); For instance, if we speak of four iterations, we actually mean eight half-iterations. The Turbo product decoder structure for a half-iteration is shown in Fig. 3, where the extrinsic information at the *m*th half-iteration is stored in matrix W(m), whereas the reliability information of the original received data is stored in matrix Λ_R . For TPC decoding, usually six to eight half-iterations are sufficient to obtain good performance results.

The reason of scaling of the extrinsic information in step (4) in the preceding algorithm is to make the reliability and weight factors independent of the chosen product code. In fact, these factors change at each halfiteration and should be optimized for the employed product code and the channel characteristics of the communications system. If scaling is employed, one



Figure 3. TPC decoder (half-iteration).

Table 1. Weight and Reliability Factors

т	0	1	2	3	4	5	6	7
lpha eta	$\begin{array}{c} 0.0 \\ 0.2 \end{array}$	$\begin{array}{c} 0.5 \\ 0.3 \end{array}$	$\begin{array}{c} 0.7 \\ 0.5 \end{array}$	$0.9 \\ 0.7$	1.0 0.9	1.0 1.0	$\begin{array}{c} 1.0\\ 1.0\end{array}$	$1.0 \\ 1.0$

suggestion for these factors versus m (number of halfiteration), is as given in Table 1 [22].

3. OPTICAL CDMA SYSTEMS

In the following sections, we describe optical orthogonal codes and optical CDMA systems using different modulation types.

3.1. Optical Orthogonal Codes (OOCs)

In an optical CDMA system, a unique OOC [3] sequence is assigned to each user; thus, each user has a predetermined address sequence. To ensure orthogonality, OOC sequences have to satisfy the following two properties [3,10]:

- 1. Each sequence should be distinguished from a timeshifted version of itself (autocorrelation constraint),
- 2. Each sequence should be distinguished from a possibly time-shifted version of another sequence (cross-correlation constraint).

The parameters of an OOC are denoted by F, K, λ_a , and λ_c , which represent codelength (i.e., number of chips), code weight, autocorrelation constraint, and crosscorrelation constraint, respectively. There exist various design techniques for OOCs [10–14]. OOCs with $\lambda_a = \lambda_c =$ 1 have the best achievable correlation constraints for a direct-detection optical CDMA system. For example, three OOC sequences are shown in Fig. 4 [15].

Here, the OOC parameters are F = 28, K = 3, and $\lambda_a = \lambda_c = 1$, and the number of users is N = 3. The sequences are denoted as $S_u = s_{u,0} \ s_{u,1} \ s_{u,2} \ \cdots \ s_{u,n_{chip}-1}$, where $u = 1, 2, \ldots, N$, $s_{u,j} \in \{0, 1\}$ and n_{chip} denotes the number of chips (i.e., F, the number of time slots in one bit interval for OOK and half the number of time slots for binary PPM). It is assumed that the K light pulses of an OOC sequence have a normalized amplitude equal to 1 and since usually $K \ll n_{chip}$, it is more convenient to indicate the OOC sequences with the chip numbers at which pulses are placed. For the sequences shown in Fig. 4, the notation would be $S_1 = \{0, 2, 8\}, S_2 = \{0, 3, 10\},$ and $S_3 = \{0, 4, 13\}$.

TURBO PRODUCT CODES FOR OPTICAL CDMA SYSTEMS 2731



Figure 4. (28,3,1,1) optical orthogonal code sequences.

As stated before, OOCs need to satisfy certain correlation constraints to ensure orthogonality. These constraints can be expressed as

1. Autocorrelation property:

$$\Theta_{uu}(m) = \sum_{i=0}^{n_{\rm chip}-1} s_{u,i} s_{u,(i+m \bmod n_{\rm chip})} \le \lambda_a$$
(14)

for any sequence S_u and any integer $m \neq 0$. $\Theta_{uu}(0) = K$ since each sequence has K pulses.

2. Cross-correlation property:

$$\Theta_{uw}(m) = \sum_{i=0}^{n_{\rm chip}-1} s_{u,i} s_{w,(i+m \bmod n_{\rm chip})} \le \lambda_c$$
(15)

for any two sequences S_u and S_w ($S_u \neq S_w$) and any integer m.

For an $(F, K, \lambda_a = 1, \lambda_c = 1)$ OOC, it can be shown that the upper bound on the number of address sequences (i.e., number of users) is equal to [3,10]

$$N_{\max} = \left\lfloor \frac{F-1}{K(K-1)} \right\rfloor \tag{16}$$

where $\lfloor x \rfloor$ denotes the integer part of the real number *x*. Perfect optimal OOCs are OOCs with length

$$F = N_{\max}K(K-1) + 1$$
 (17)

3.2. System Model

Figure 5 shows the block diagram of an intensity modulated, direct-detection fiberoptic CDMA network that employs all-optical signal processing. The configuration shown is a star network with N users; however, other structures like ring networks are also possible.

A user in the system shown in Fig. 5 sends data using the address sequence of the destination user. Let us now briefly describe how this is established. At the transmitter side, the binary source is followed by a modulator. The modulation type is OOK or PPM. In case of OOK, the user transmits bit 1 by sending the address sequence, whereas bit 0 is transmitted by leaving the bit interval



Figure 5. Optical CDMA system.

empty. The major drawbacks of OOK systems are that synchronization and baseline wander problems [1] may occur as a result of long sequences of zeros or ones. To avoid these, scramblers [23] or $m\mathbf{B}n\mathbf{B}$ codes [1] could be used. An alternative to OOK is PPM [6,7]. In case of binary PPM, the bit interval is divided into two time slots and binary 1 and 0 are sent as OOC sequences in their assigned time slots.

The modulator in Fig. 5 is followed by a bit-to-pulse converter to generate a short light pulse of duration T_c , which is the duration of one chip. If T_b denotes the bit duration, then for OOK, $T_c = T_b/F$, whereas for binary PPM, $T_c = T_b/(2F)$. The bandwidth expansion factor is equal to F and 2F for OOK and binary PPM, respectively. An optical switch is placed after the bit-to-pulse converter. The purpose of this switch is to select the OOC sequence of the destination user. After selecting the destination address, the OOC sequence is formed by passing the short light pulse through K fiber delay lines, which place the K pulses according to the OOC sequence of the destination user. All users are transmitting their data via an $N \times N$ optical star coupler. Hence, all users are receiving the same signals at overlapping times and wavelengths.

An optical CDMA receiver consists of the following devices: an optical hard-limiter [3], a correlator, and an OOK or PPM demodulator. The output of the hard-limiter is modeled as

$$h(\phi) = \begin{cases} 1 & \text{if } \phi \ge 1\\ 0 & \text{if } \phi < 1 \end{cases}$$
(18)

where ϕ is the normalized input light intensity. Employment of the hard-limiter at the receiver is optional. However, it is shown that placing an optical hardlimiter before the optical correlator enhances the system performance significantly [3,4]. In fact, performance can be improved further by using two hard-limiters, one before and one after the correlator [25]. This article considers the employment of a single hard-limiter or no hard-limiter at all.

Like the encoder, the correlator is also a configuration composed of fiber delay lines that are matched to the destination user's address sequence. Each receiver receives the same sequences; however, because of the orthogonality of the sequences, the data can be decoded only by its intended correlator. The correlator are followed by a demodulator after which the data are finally recovered.

3.3. Channel Model for Optical CDMA with OOK

Without loss of generality, user 1 can be considered as the destination address. Also, the main source of noise can be assumed as optical multiple access interference caused by other users. For an OOK-CDMA system with hard-limiter, the correlator output of user 1 may be modeled as

$$Z_{u_1,\text{HL}} = \begin{cases} K & \text{if } x_j = 1\\ I_{u_1,\text{HL}} & \text{if } x_j = 0 \end{cases}$$
(19)

and for an OOK system without hard-limiter

$$Z_{u_1,\text{NHL}} = x_j K + I_{u_1,\text{NHL}} \tag{20}$$

Here, $x_j \in \{0, 1\}$ represents the data sent to user 1, $I_{u_1,\text{HL}}$ is the interference signal due to other users for a system with hard-limiter, $I_{u_1,\text{NHL}}$ is the interference signal due to other users for a system without hard-limiter, and K is the weight of the OOC. In the correlator output signals, photodetector and amplifier noise may be ignored since these are relatively small if compared to optical multiple access interference.

In an OOK system, the decision criterion at the receiver is

$$y = \begin{cases} 0, & \text{if } Z_{u_1} < \text{Th} \\ 1, & \text{if } Z_{u_1} \ge \text{Th} \end{cases}$$
(21)

where $\text{Th}(0 < \text{Th} \le K)$ is a predefined threshold level; thus, if the signal level is above Th, it is assumed that bit 1 was received; otherwise, the assumption is made that the received signal represents bit 0. The probability of bit error for an OOK system may be expressed as

$$P_{e,OOK} = \Pr(r_j = 1 \mid x_j = 0) \Pr(x_j = 0) + \Pr(r_j = 0 \mid x_j = 1) \Pr(x_j = 1)$$
(22)

Since direct detection is used, we observe that for OOK, $Pr(r_j = 0 | x_j = 1) = 0$. Hence, the OOK system is characterized by an asymmetric channel called *Z* channel [5] as shown in Fig. 6.

Under the assumption that $Pr(x_j = 1) = Pr(x_j = 0) = \frac{1}{2}$, we can write the conditional probability $Pr(r_j | x_j)$ as

$$\Pr(r_j \mid x_j) = \begin{cases} 2P_{e,OOK} & \text{if } r_j = 1, x_j = 0\\ 1 - 2P_{e,OOK} & \text{if } r_j = 0, x_j = 0\\ 1 & \text{if } r_j = 1, x_j = 1\\ 0 & \text{if } r_j = 0, x_j = 1 \end{cases}$$
(23)

It is shown that for an OOK system with hard-limiter, the probability of bit error can be upper-bounded by [3]

$$P_{e,\text{OOK,HL}} \le \frac{1}{2} \binom{K}{\text{Th}} \prod_{i=1}^{\text{Th}} (1 - v^{N-i})$$
(24)

where N is the number of active users and

$$v = 1 - \frac{K}{2F} \tag{25}$$



Figure 6. Z-channel model.

For an OOK system without hard-limiter, the probability of bit error has the following upperbound [3]:

$$P_{e,\text{OOK,NHL}} \le \frac{1}{2} \sum_{i=\text{Th}}^{N-1} \binom{N-1}{i} q^i (1-q)^{N-1-i}$$
(26)

where

$$q = \frac{K^2}{2F} \tag{27}$$

3.4. Channel Model for Optical CDMA with Binary PPM

As indicated earlier, for an optical CDMA system with binary PPM, one bit duration is split into two time slots. We assume that bit 0 is transmitted as an OOC sequence in slot Z_0 and that bit 1 is transmitted as an OOC sequence in slot Z_1 . Let Z_{0,u_1} and Z_{1,u_1} denote the correlator outputs for the two time slots of user 1. At the receiver, the decision as to whether bit 1 or 0 is received is made by

$$r_{j} = \begin{cases} 0 & \text{if } Z_{0,u_{1}} > Z_{1,u_{1}} \\ 1 & \text{if } Z_{1,u_{1}} > Z_{0,u_{1}} \end{cases}$$
(28)

In case of equality, $Z_{1,u_1} = Z_{0,u_1}$, one bit is randomly chosen. PPM systems have the main advantage that there is no need to define a threshold Th as in the case of OOK. The drawback is that PPM requires more system bandwidth; binary PPM occupies twice as much bandwidth as does OOK.

For binary PPM, the probability of bit error, $P_{e,\text{PPM}}$, is equal to

$$P_{e,\text{PPM}} = \Pr(Z_{0,u_1} \ge Z_{1,u_1} \mid x_j = 1) \Pr(x_j = 1) + \Pr(Z_{1,u_1} \ge Z_{0,u_1} \mid x_j = 0) \Pr(x_j = 0)$$
(29)

Assuming $Pr(x_j = 1) = Pr(x_j = 0) = \frac{1}{2}$, an optical PPM-CDMA system can be characterized as a binary symmetric channel (BSC) as shown in Fig. 7, where, the crossover probabilities are the same. The transition probabilities can be expressed as

$$\Pr(r_j \mid x_j) = \begin{cases} P_{e,\text{PPM}} & \text{if } r_j = 1, x_j = 0\\ 1 - P_{e,\text{PPM}} & \text{if } r_j = 0, x_j = 0\\ 1 - P_{e,\text{PPM}} & \text{if } r_j = 1, x_j = 1\\ P_{e,\text{PPM}} & \text{if } r_j = 0, x_j = 1 \end{cases}$$
(30)



$Pr(r_j = 1 x_j = 0) = Pr(r_j = 0 x_j = 1)$	
$\Pr(r_j = 1 x_j = 1) = \Pr(r_j = 0 x_j = 0)$	

Figure 7. Binary symmetric channel.

For an optical PPM system without hard-limiter, the correlator outputs for the two time slots for user 1 can be written as

$$Z_{0,u_1,\text{NHL}} = (1 - x_j)K + I_{0,u_1,\text{NHL}}$$
(31)

and

$$Z_{1,u_1,\text{NHL}} = x_j K + I_{1,u_1,\text{NHL}}$$
(32)

where $x_j \in \{0, 1\}$ represents the transmitted bit to user 1; $I_{0,u_1,\text{NHL}}$ is the interference due to other users sending bit 0 and similarly, $I_{1,u_1,\text{NHL}}$ is the interference due to other users sending bit 1. Using (29) and following the analysis in Ref. 16, the probability of bit error for an optical PPM-CDMA system without hard-limiter can be obtained as

$$P_{e,\text{PPM,NHL}} = \sum_{i=K}^{N-1} \sum_{m=0}^{N-1-i} \binom{N-1}{m} \binom{N-1}{m+i} \times q^{2m+i} (1-q)^{2(N-1-m)-i}$$
(33)

If we consider an optical PPM-CDMA system with hardlimiter, the correlator outputs for the two time slots for user 1 can be written as

$$Z_{0,u_1,\text{HL}} = \begin{cases} K & \text{if } x_j = 0\\ I_{0,u_1,\text{HL}} & \text{if } x_j = 1 \end{cases}$$
(34)

and

$$Z_{1,u_1,\text{HL}} = \begin{cases} K & \text{if } x_j = 1\\ I_{1,u_1,\text{HL}} & \text{if } x_j = 0 \end{cases}$$
(35)

where $x_j \in \{0, 1\}$ represents the transmitted bit to user 1; $I_{0,u_1,\text{HL}}$ is the interference due to other users sending bit 0 and similarly, $I_{1,u_1,\text{HL}}$ is the interference due to other users sending bit 1. Applying (29) and assuming equally likely symbols x_j , we obtain the probability of bit error for an optical PPM-CDMA system with hard-limiter as [16]

$$P_{e,\text{PPM,HL}} = \prod_{i=1}^{K} (1 - v^{N-i})$$
(36)

Until now, we presented the probability of bit errors and transition probabilities for four possible cases of optical CDMA systems, namely, OOK with and without hardlimiter and binary PPM with and without hard-limiter. We observed that an OOK system is a binary asymmetric memoryless channel, whereas a binary PPM system is a binary symmetric memoryless channel. How to implement turbo product codes in these systems is described next.

4. APPLICATION OF TURBO PRODUCT CODES IN OPTICAL CDMA

To increase the performance (and hence, the number of users) of an optical CDMA system, a TPC encoder may be



Figure 8. Optical CDMA transmitter with TPC encoder.

employed right before the modulator at the transmitter side in Fig. 8.

As shown, all data is encoded as a product code before it is modulated and passed to the bit-to-pulse converter. In the previous section, upper bounds were presented for the probability of bit errors and transition probabilities for the binary memoryless channels which characterize optical CDMA systems. These results are summarized in Tables 2 and 3. With this information, the implementation of the turbo product decoding algorithm described earlier is possible.

From our discussion on turbo product decoding, we know that initial reliability information is needed on received symbol r_j in order to apply the iterative decoding procedure. For an OOK system, we find the reliability

Table 2. Probability of Bit Error for Optical CDMA Systems

P_e	Upperbound
$P_{e,\mathrm{OOK,HL}}$	$rac{1}{2} {K \choose Th} \prod_{i=1}^{ ext{Th}} (1-v^{N-i})$
$P_{e,\mathrm{OOK,NHL}}$	$rac{1}{2}\sum_{i=Th}^{N-1} \binom{N-1}{i} q^i (1-q)^{N-1-i}$
$P_{e,\mathrm{PPM,HL}}$	$\prod_{i=1}^{K} (1-v^{N-i})$
$P_{e,\mathrm{PPM,NHL}}$	$\sum_{i=K}^{N-1} \sum_{m=0}^{N-1-i} \binom{N-1}{m} \binom{N-1}{m+i} q^{2m+i} (1-q)^{2(N-1-m)-i}$

information for r_j as

$$\Lambda(r_j = 1) = \log\left(\frac{1}{2P_{e,\text{OOK}}}\right) \tag{37}$$

and

$$\Lambda(r_j = 0) = -\infty \tag{38}$$

where $P_{e,OOK}$ is upper-bounded by (24) or (25) depending on whether a hard-limiter is used or not. The interpretation of (37) and (38) is that we are sure about a received 0, whereas a received 1 could be in error and has the reliability given in (37). Hence, when applying the Turbo product decoding algorithm in a direct detection OOK-CDMA system, only those bit positions that have initially (37) as their reliability information need to be considered when evaluating the extrinsic information of the received product codeword.

For the binary PPM system, the initial reliability information of r_i is obtained by

$$\Lambda(r_{j}) = \begin{cases} \log\left(\frac{1-P_{e,\text{PPM}}}{P_{e,\text{PPM}}}\right) & \text{if } r_{j} = 1\\ \log\left(\frac{P_{e,\text{PPM}}}{1-P_{e,\text{PPM}}}\right) & \text{if } r_{j} = 0 \end{cases}$$
(39)

Since both symbols are equally likely, reliability and extrinsic information must be obtained for all bit positions when applying TPCs to optical PPM-CDMA systems.

We observe that in both OOK and PPM systems, the number of active users (i.e., N) is required to estimate the probability of bit error P_e . This information could be retrieved using an estimator placed at the front end at the receiving side (see Fig. 9).

Depending on the total light power per bit period, this device can estimate the number of active users and pass this information to the TPC decoder, which is then capable of calculating an upper bound of P_e depending on the modulation type and whether an optical hard-limiter is employed. The TPC decoder could also perform a lookup

Table 3. Transition Probabilities						
System	$\Pr(r_j = 0 \mid x_j = 0)$	$\Pr(r_j = 1 \mid x_j = 0)$	$\Pr(r_j = 0 \mid x_j = 1)$	$\Pr(r_j = 1 \mid x_j = 1)$		
OOK PPM	$1-2P_{e,\mathrm{OOK}}\ 1-P_{e,\mathrm{PPM}}$	${2P_{e,{ m OOK}} \over P_{e,{ m PPM}}}$	$0 P_{e,\mathrm{PPM}}$	$1 \ 1 - P_{e,\mathrm{PPM}}$		



Figure 9. Optical CDMA receiver with TPC decoder.

if a table for P_e versus N is stored beforehand. Having an estimate of P_e enables the TPC decoder to perform iterative decoding via calculation of extrinsic and initial reliability information as described before.

5. PERFORMANCE RESULTS

The following simulations were carried out in another study by the authors [16] to demonstrate the performance improvement due to the employment of TPCs in optical CDMA systems. It is assumed that optical multiple access interference is the main source of noise. Other noise sources, such as photodetector and thermal noise and losses related to optical transmission, are ignored. Transmission of all OOC sequences is modeled to be chip-synchronous, which is a worst-case assumption [3]. Furthermore, it is assumed that perfect optimal OOCs are employed.

The first set of simulations presented here is carried out for an optical OOK-CDMA system with hard-limiter. Assuming that optical multiple access interference is the main source of interference, the threshold at the receiver side is set to Th = K - 1. Figure 10 shows the bit error rate (BER) versus number of active users for various coded and uncoded systems.

The BER curves for the uncoded systems with OOCs of weight K = 4, K = 5, and K = 6 are plotted using the upperbound for $P_{e,OOK,HL}$. It is observed that a K = 4 system with hard decision and a BCH(64,57,4) linear block code or a Reed–Solomon RS(255,239,17) linear block code shows minor performance improvement if compared to the uncoded K = 4 system. On the other hand, a K = 4 system with a TPC BCH(64,57,4)² code (i.e., both component codes of the product code are BCH(64,57,4) codes) outperforms the previously mentioned systems,

the uncoded K = 5 system, and even the uncoded K = 6 system. The number of half-iterations for the TPC is 6, and the *f* parameter for the Chase algorithm is chosen to be 4 (i.e., 16 test patterns are formed for each row or column of the product codeword). The TPC decoder uses the weight and reliability factors given in Table 1. Similar results as those observed in Fig. 10 are obtained for an optical OOK-CDMA system with no hard-limiter and the same conditions as above.

The performance curves for an optical PPM-CDMA system with no hard-limiter is shown in Fig. 11. Again, the TPC BCH $(64,57,4)^2$ coded system outperforms all other coded and uncoded systems. It has been shown [16] that similar performance improvements may be achieved for optical PPM-CDMA systems with hard-limiter.

All simulation results show that for a given BER, it is possible to implement optical CDMA systems with lowweight OOCs (i.e., low K) if a TPC is employed. The performance curves show that we might switch from an uncoded K = 6 system to a TPC coded K = 4 system and obtain an improved BER. The TPC discussed here has an coding overhead of 25%; nevertheless, if we compare the K = 4 system with the K = 6 systems (the ratio of OOC codelengths vs. number of users as shown in Fig. 12), we observe that the K = 6 system requires OOCs with about 2.5 times more codelength.

Considering the coding overhead, the net reduction in bandwidth expansion factor is $2\times$ for the system under consideration. Hence, for a given BER, the employment of a TPC reduces the required OOC codelength significantly and enables an increase in available bandwidth. Furthermore, having a smaller *K* value means fewer fiber delay lines in the encoders and correlators. A reduction in *K* provides also a reduction in the amount of power to be transmitted because the



Figure 10. Performance of optical OOK-CDMA system with hard-limiter.



Figure 11. Performance of optical PPM-CDMA system without hard-limiter.

attenuation of the light pulses decreases when they are split into fewer fiber delay lines. Using a TPC BCH(64,57,4)² code requires about 1 dB more power to be transmitted per bit. However, switching from K = 6 to K = 4 provides a gain of 1.76 dB. As a result, for the TPC coded system, the net power gain would be 0.76 dB.

6. CONCLUSION

The wide deployment of optical CDMA systems was not possible in the past because of the huge bandwidth expansion imposed by very long optical orthogonal code (OOC) sequences. However, it appears that using error correction codes might be an efficient way to reduce bandwidth expansion for a given BER target, enabling wider application possibilities of optical CDMA systems.



Figure 12. OOC codelength ratio $F_{(K=6)}/F_{(K=4)}$.

In this article, the application of Turbo product codes (TPC) was considered for an intensity modulated, directdetection optical CDMA system with OOK or binary PPM. It was shown that TPC-coded optical CDMA systems have the potential to outperform their uncoded counterparts or systems with simple one-dimensional error correction codes. Furthermore, it was observed that the usage of OOC sequences with less weight and length is possible via TPCs. TPCs discussed here use only hard (thresholded) decisions and estimates of the number of active users in the system, which make them much easier to fit into existing networks.

Beside optical CDMA systems, other interesting approaches would be the application of TPCs in optical systems with dense wavelength division multiplexing (DWDM) [26]. Also, the application in hybrid WDMA/CDMA or TDMA/CDMA systems [24] might be considered.

BIOGRAPHIES

Steven W. McLaughlin received his B.S.E.E. degree in 1985 from Northeastern University, Boston, Massachusetts, his M.S.E. degree in 1986 from Princeton University, New Jersey, and a Ph.D. degree in 1992 from the University of Michigan. He joined the Electrical Engineering Department at the Rochester Institute of Technology, New York, in 1992. In 1996, he joined the School of ECE at the Georgia Institute of Technology, Atlanta, where he is now associate professor. He holds 21 U.S. patents and has more than a dozen pending in the areas of coding and signal processing for data storage and fiber optic transmission systems. In 1997, he received the Presidential Early Career Award for Scientists and Engineers PECASE and was cited by President Clinton for "for leadership in the development of high-capacity, nonbinary optical recording formats." He also received the National Science Foundation CAREER Award in 1997. His areas of interest are communication and information theory with specific interest in coding for data storage and fiber optic transmission systems.

Cenk Argon received his B.S.E.E. degree in 1992 and his M.S.E.E. degree in 1994 from the Middle East Technical University, Ankara, Turkey, and his M.S. degree in 1999 from the Georgia Institute of Technology, Atlanta. He is currently pursuing his Ph.D. degree and working as a graduate research assistant in the Communications and Information Theory Research Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology. His areas of interest are forward error correction (FEC) and signal processing techniques for optical networks, turbo product codes, and optical CDMA systems.

BIBLIOGRAPHY

- 1. G. Keiser, *Optical Fiber Communications*, 3rd ed., McGraw-Hill, 2000.
- 2. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, 1995.
- J. A. Salehi, Code division multiple access techniques in optical fiber networks—Part I: Fundamental principles, *IEEE Trans. Commun.* 37: 824–833 (Aug. 1989).
- 4. J. A. Salehi and C. A. Brackett, Code division multiple access techniques in optical fiber networks—Part II: Systems performance analysis, *IEEE Trans. Commun.* **37**: 834–842 (Aug. 1989).
- 5. J.-G. Zhang, Use of error-correction codes to improve the performance of optical fiber code division multiple access systems, *Proc. URSI Int. Symp. Signals, Systems, and Electronics*, 1998, pp. 361–366.
- J. Y. Kim and H. V. Poor, Turbo-coded optical direct-detection CDMA system with PPM modulation, *IEEE J. Lightwave Technol.* 19(3): 312–323 (March 2001).
- T. Ohtsuki and J. M. Kahn, BER performance of turbo-coded PPM CDMA systems on optical fiber, *IEEE J. Lightwave Technol.* 18(12): 1776–1784 (Dec. 2000).
- 8. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. IEEE ICC*, 1993, pp. 1064–1070.
- C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* 44(10): 1261-1271 (Oct. 1996).
- F. R. K. Chung, J. A. Salehi, and V. K. Wei, Optical orthogonal codes: Design, analysis, and applications, *IEEE Trans. Inform. Theory* 35: 595-604 (May 1989).
- A. A. Shaar and P. A. Davies, Prime sequences: Quasioptimal sequences for OR channel code division multiplexing, *Electron. Lett.* 19(21): 888–890 (Oct. 1983).
- S. V. Maric, Z. I. Kostic, and E. L. Titlebaum, A new family of optical code sequences for use in spread spectrum fiber-optic local area networks, *IEEE Trans. Commun.* 41(8): 1217–1221 (Aug. 1993).
- H. Chung and P. V. Kumar, Optical orthogonal codes-New bounds and an optimal construction, *IEEE Trans. Inform. Theory* 36(4): 866-873 (July 1990).

- C. Argon and R. Ergül, Optical CDMA via shortened optical orthogonal codes based on extended sets, *Opt. Commun.* 116: 326–330 (May 1995).
- C. Argon and R. Ergül, Detection of shortened OOC codewords in optical CDMA systems with double hardlimiters, Opt. Commun. 177: 277-281 (April 2000).
- C. Argon and S. W. McLaughlin, Optical OOK-CDMA and PPM-CDMA systems with turbo product codes, *IEEE J. Lightwave Technol.* (in press).
- R. Pyndiah, Near-optimum decoding of product codes: Block turbo codes, *IEEE Trans. Commun.* 46(8): 1003–1010 (Aug. 1998).
- P. Elias, Error-free coding, *IRE Trans. Inform. Theory* IT-4: 29–37 (Sept. 1954).
- J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* 42: 429-445 (March 1996).
- D. Chase, A class of algorithms for decoding block codes with channel measurement information, *IEEE Trans. Inform. Theory* **IT-18**(1): 170–182 (Jan. 1972).
- 21. S. B. Wicker, Error Control Systems for Digital Communication and Storage, Prentice-Hall, 1995.
- 22. A. Picart and R. Pyndiah, Adapted iterative decoding of product codes, *Proc. IEEE Globecom*, 1999, pp. 2357–2362.
- B. P. Lathi, Modern Digital and Analog Communication Systems, 2nd ed., Holt, Rinehart and Winston, 1989.
- 24. W. Huang, M. H. M. Nizam, I. Andonovic, and M. Tur, Coherent optical CDMA (OCDMA) systems used for highcapacity optical fiber networks—system description, OTDMA comparison and OCDMA/WDMA networking, *IEEE J. Lightwave Technol.* 18: 765–778 (June 2000).
- T. Ohtsuki, Performance analysis of direct-detection optical asynchronous CDMA systems with double optical hardlimiters, *IEEE J. Lightwave Technol.* 15(3): 452–457 (March 1997).
- O. Aitsab and V. Lemaire, Block turbo code performances for long-haul DWDM optical transmission systems, *Proc. OFC*, 2000, pp. 280-282.

TURBO TRELLIS-CODED MODULATION (TTCM) EMPLOYING PARITY BIT PUNCTURING AND PARALLEL CONCATENATION

PATRICK ROBERTSON Institute for Communications Technology German Aerospace Center (DLR) Wessling, Germany

THOMAS WÖRZ Audens ACT Consulting GmbH Wessling, Germany

1. INTRODUCTION

In 1993, powerful Turbo codes were introduced [1] that achieve good bit error rates $(10^{-3} \cdots 10^{-5})$ at very low SNR close to the channel capacity. They are of interest in a wide

range of digital telecommunications applications such as satellite communications and mobile radio. For instance, they are deployed in the air interface of the UMTS mobile radio system [2]. They comprise two binary component codes and an interleaver and were originally proposed for binary modulation schemes [e.g., BPSK (binary phase shift keying)]. Successful attempts were soon undertaken to combine binary Turbo codes with higher-order modulation [e.g., 8-PSK, 16-QAM (quadrature amplitude modulation)] using Gray mapping [3], and alternatively as component codes within multilevel codes [4]. In contrast, in the approach called Turbo trellis-coded modulation (TTCM) one employs two Ungerboeck type codes [5] in combination with trellis-coded modulation (TCM) in their recursive systematic form as component codes in an overall structure rather similar to binary Turbo codes [6].

Before the combination of Turbo codes and TCM is illustrated, the main idea behind conventional TCM codes is briefly reviewed. Generally, the bit error rate (BER) of an uncoded modulation scheme, at least for medium to large signal-to-noise ratio E_s/N_0 , depends on the minimum squared Euclidean distance between all members of the considered signal set (e.g., QPSK: $d_{\text{free}}^2 = 4$). Increasing $d_{\rm free}^2$ can be achieved by standard channel coding, which adds coded bits to the information bits. As a result, more bandwidth is required to transmit the information bits including the redundancy with the same signal set. However, in the case of TCM the signal set is basically extended such that the same bandwidth is occupied compared to the uncoded case, for example, from QPSK to 8-PSK. Then, a rate- $\frac{2}{3}$ convolutional code is applied to generate sequences of 8-PSK symbols, whose $d_{
m free}^2$ is larger than for the uncoded QPSK transmission. The design of the convolutional code is based on the set partitioning of the underlying signal set aiming at the optimization of d_{free}^2 . Summarizing, one transmits the same number of information bits per modulation symbol with increased $d_{\rm free}^2$ and without increasing the necessary bandwidth. The principle is applicable in the same way to higherorder modulation schemes such as 16- and 64-QAM. Turbo trellis-coded modulation codes can be decoded with the Viterbi or the Bahl-Jelinek (symbol-by-symbol MAP) algorithm [7]. Multidimensional TCM allows even higher bandwidth efficiency than traditional Ungerboeck TCM by assigning more than one symbol per trellis transition or step [8]. In this case, the set partitioning takes into account the union of more than one two-dimensional signal set.

The basic principle of Turbo codes is applied to TCM by retaining the important properties and advantages of both their structures. Essentially, TCM codes can be seen as systematic feedback convolutional codes followed by one (or more for multidimensional codes [8]) signal mapper(s). Just as binary Turbo codes use a parallel concatenation of two binary recursive convolutional encoders, in TTCM one concatenates two recursive TCM encoders, and adapts the interleaving and puncturing. Naturally, this has consequences at the decoding side, which are explained in depth later.

One can also apply the concept of TTCM to incorporate multidimensional component codes, which allows a higher overall bandwidth efficiency for a given signal constellations than ordinary TTCM. By applying the technique to 8-PSK, 16-QAM, and 64-QAM modulation formats, we will show its viability over a large range of bandwidth efficiency and signal-to-noise ratios. In all cases, low BERs $(10^{-4} \cdots 10^{-5})$ can be achieved within 1 dB or less from Shannon's limit — a finding that in the context of binary Turbo codes was responsible for the interest they generated.

The article begins by describing the generic encoder (beginning with a motivation for its structure); an encoder with 8-PSK signaling will serve as a salient example. We then present the results of a search for component codes, taking into consideration the puncturing at the encoder. This is followed by a section on the iterative decoder using symbol-by-symbol MAP component decoders whose structures are derived for our case of nonbinary trellises and special metric calculation. Finally, we present simulation results of the TTCM scheme with two- and four-dimensional 8-PSK, as well as two-dimensional 16-QAM and 64-QAM. The influence of varying the block size and interleaver type-both of important practical relevance-is also subject of investigation. We conclude by presenting further literature to the subject and current areas of investigation.

2. THE ENCODER

2.1. Motivation for the Structure

Let us recall that two important characteristics of Turbo codes are their simple use of recursive systematic component codes in a parallel concatenation scheme. Pseudorandom bitwise interleaving between encoders ensures a small bit error probability [9]. What is crucial to their practical suitability is the fact that they can be decoded iteratively with good performance [1]. It is well known that Ungerboeck codes combine coding and modulation by optimizing the Euclidean distance between codewords and achieve high spectral efficiency (m bits)per 2^{m+1} -ary symbol from the two-dimensional signal space) through signal set expansion. The encoder can be represented as combination of a systematic recursive convolutional encoder and symbol mapper. If \tilde{m} out of m bits are encoded, the resulting trellis diagram consists of $2^{\tilde{m}}$ branches per state, not counting parallel transitions. This results in more than two branches per state for $\tilde{m} > 1$ —we call this a *nonbinary trellis*.

We have employed Ungerboeck codes (and multidimensional TCM codes) as building blocks in a Turbo coding scheme in a similar way as binary codes were used through so-called parallel concatenation [1]. The major differences are (1) the interleaving now operates on short groups of m bits (e.g., pairs for 8-PSK with twodimensional TCM schemes) instead of single bits; (2) to achieve the desired spectral efficiency, puncturing the parity information is not quite as straightforward as in the binary Turbo coding case; and (3) there are special constraints on both the component encoders as well as the structure of the interleaver.

2.2. Definition of the Encoder

In this section we begin by defining the generic encoder for TTCM and then continue to illustrate a simple



Figure 1. The generic encoder that treats uncoded bits as coded bits from a structural point of view.

example encoder. Figure 1 shows this generic encoder, comprising two TCM encoders linked by the interleaver. It is important to remember that the interleaver operates on small groups of *m* bits. Let the size of the interleaver — the number of these groups — be N. The number of modulated symbols per block is $N \cdot n$, with n = D/2, where D is the signal set dimensionality. The number of information bits transmitted per block is $N \cdot m$. The encoder is clocked in steps of $n \cdot T$ where T is the symbol duration of each transmitted $2^{(m+1)/n}$ -ary symbol. In each step, m information bits are input and n symbols are transmitted, yielding a spectral efficiency of m/n bits per symbol usage. A signal mapper follows each recursive systematic convolutional encoder where the latter each produce one parity bit in addition to retaining the *m* information bits at their inputs. For clarity we have not depicted any special treatment of the $m - \tilde{m}$ uncoded bits as opposed to the \tilde{m} bits to be encoded: in practice, uncoded bits would not need to be passed through the interleaver but would simply be used to choose the final signal point from a subset of points after the selector. We will return to the problem of parallel transitions shortly.

For the moment, the interleaver is restricted to keeping each group of m bits unchanged within itself (as visualized by the dashed lines passing through the interleaver in Fig. 1). The output of the lower encoder/mapper is deinterleaved according to the inverse operation of the interleaver. This ensures that at the input of the selector, the m information bits partly defining each group of nsymbols of both the upper and lower input are identical. Therefore, if the selector is switched such that a group of n symbols is chosen alternately from the upper and lower inputs, then the sequence of $N \cdot n$ symbols at the output has the important property that each of the N groups of m information bits defines part of each group of n output symbols. The remaining bit, which is needed to define each group of n symbols, is the parity bit taken alternatively from the upper and lower encoders.

A simple example will now serve to clarify the operation of the encoder for the case n = 1, m = 2, N = 6 and 8-PSK signaling: it is illustrated in Fig. 2. The set partitioning is shown in Fig. 3. The 6-long sequence $(d_1, d_2, \ldots, d_6) = (00, 01, 11, 10, 00, 11)$ of information bit pairs (m = 2) is encoded in an Ungerboeck style encoder to yield the 8-PSK sequence (0,2,7,5,1,6). The information bits are interleaved-on a pairwise basis- and encoded again into the sequence (6,7,0,3,0,4). We deinterleave the second encoder's output symbols to ensure that the ordering of the two information bits partly defining each symbol corresponds to that of the 1st encoder; thus we now have the sequence (0,3,6,4,0,7). Finally, we transmit the 1st symbol of the first encoder, the second symbol of the second encoder, the third of the first encoder, the fourth symbol of the second encoder, and so on: (0,3,7,4,1,7). Thus the parity bit is alternately chosen from encoders 1 and 2 (bold, non-bold, bold \ldots). Also, the kth information bit pair exactly determines 2 of the 3 bits of the *k*th symbol x_k . This ensures that each information bit pair defines part of the constellation of an 8-PSK symbol exactly once.



Figure 2. The encoder shown for 8-PSK with two-dimensional component codes memory 3. An example of interleaving with N = 6 is shown. Bold letters indicate that symbols or pairs of bits correspond to the upper encoder.

2.3. Interleaver and Code Constraints

2.3.1. Basic Interleaver Types. By deinterleaving the output of the second encoder, each symbol index kbefore the selector in Fig. 1 has the property of being associated with input information bit group index k, regardless of the actual interleaving rule. However, to ensure that punctured and unpunctured symbols are uniformly spread, that is, occur alternately, at the input of both decoders, the interleaver must map even positions to even positions and odd ones to odd ones (or even-odd, odd-even). Other than this constraint, the interleaver can be chosen to be pseudorandom or modified to avoid low distance error events. It is important to remember that we have so far assumed that the interleaver keeps the input unchanged within each group of information bits and that the corresponding symbol deinterleaver does not modify its symbol inputs (except for the actual reordering of their positions, of course); see the top example in Fig. 4.

We have also forced a constraint on the component code such that the corresponding trellis diagram of the convolutional encoders should have no parallel transitions. This ensures that each information bit benefits from the parallel concatenation and interleaving. This condition can be relaxed in a number of cases. The first [10] applies if the interleaver no longer keeps each group of m bits unchanged during interleaving: This method alters the position of the bits within each group as they are interleaved, see the bottom example in Fig. 4. The argument is that each information bit should influence the state of at least one encoder; bits that lead to only parallel transitions in *one* encoder will thus cause the *other* encoder to change its state and accrue distance henceforth. All bits would thus benefit from the interleaving and parallel concatenation. In one study [10] a slight performance gain was reported for the first few iterations compared to TTCM schemes with no parallel transitions.

The second case in which we allow parallel transitions in the component code is when we desire a very high bandwidth efficiency. Because of the higher operating SNR and the large Euclidean distance that separates the subsets of signal points that define parallel transitions (e.g., the lowest partition step in Fig. 3), corresponding uncoded information bits receive ample protection in cases such as 8-PSK transmitting 2.5 information bits per symbol and 64-QAM with 5 bits per symbol, which we will investigate below. The transmission of uncoded bits has been proposed for the multilevel approach [4], where channel capacity arguments show that when 5 information bits are sent using one 64-QAM symbol, the last 2 partitioning bits theoretically need only minimal (if any) coding protection.

2.3.2. Design Rule for Selecting the Number of Uncoded Bits. In the following, a heuristic rule is given in order to determine the number of uncoded bits per symbol. It is based on the experience that the BER of TTCM schemes (with large block lengths) reaches a value of $P_b \approx 10^{-5}$ at a signal-to-noise ratio E_s/N_0^* , which is approximately 1 dB above the corresponding channel capacity. Let us consider the sequence of increasing inner-set distances Δ_i when following down the partitioning of the corresponding signal set (for an example of partitioning an 8-PSK constellation,



Figure 3. The set partitioning for 8-PSK. Dotted ovals denote subsets corresponding to the different combinations of *d*. The distances Δ_i are relevant for code design.



Figure 4. Two kinds of interleaver for Turbo TCM, block length N = 3. *Top*:—position invariant group interleaving; *bottom*:—position swapping group interleaving.

refer to Fig. 3). For each distance we can evaluate a rough approximation of the BER in the uncoded case, by applying the well-known formula [11]

$$P_b(\Delta_i) = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_s \Delta_i}{4N_0}}$$
(1)

By using this formula to approximate the BER of the uncoded bits with $P_b(\Delta_i)$, two approximations are included:

- The error propagation from the partition levels that include coded bits into the partition levels with uncoded bits is neglected.
- Moreover, the number of nearest neighbors is not included in the calculation, only the pure distance is used to evaluate (1).

As a result, we can identify at which level of the partition chain the corresponding uncoded bits have enough protection based on the distance Δ_i and the given signal-to-noise ratio E_s/N_0^* to bring the BER below $P_b = 10^{-5}$. Two examples are given in the following:

• Example 1:

Signal set: four-dimensional 8-PSK. Desired information rate: 2.5 bits per symbol. The two 8-PSK symbols are generated by the following rule [8]: $\binom{y_1}{y_2} = z^5\binom{4}{4} + z^4\binom{0}{4} + z^3\binom{2}{2} + z^2\binom{0}{2} + z^1\binom{1}{1} + z^0\binom{0}{1}$ modulo 8. The parity bit is z^0 ; the information bits are z^1 to z^5 . Corresponding channel capacity: 8.8 dB [5] $\Rightarrow E_s/N_0^* = 9.8$ dB.

Sequence of distances Δ_i for the partition chain of the signal set [8] and corresponding uncoded bit error rates

Partition Level	Δ_i	$P_b(\Delta_i)$
0	0.586	0.05
1	1.172	0.009
2	2	0.001
3	4	$6.5\cdot 10^{-6}$
4	4	$6.5\cdot 10^{-6}$
5	8	$3.5\cdot10^{-10}$
6	∞	

- Conclusion: 3 encoded bits (including the parity bit) are necessary to reach the desired BER for the uncoded bits (hence $\tilde{m} = 2$).
- Example 2:

Signal set: two-dimensional 64-QAM.

Desired information rate: 5 bits per symbol.

Corresponding channel capacity: 16.2 dB [5] $\Rightarrow E_s/N_0^* = 17.2$ dB.

Sequence of distances Δ_i for the partition chain of the signal set [5] and corresponding uncoded bit error rates

Partition Level	Δ_i	$P_b(\Delta_i)$
0	0.095	0.06
1	0.19	0.013
2	0.38	$8\cdot 10^{-4}$
3	0.76	$4\cdot 10^{-6}$
4	1.52	$1.3\cdot10^{-10}$
5	3.05	$1.8\cdot10^{-19}$
6	∞	_

Conclusion: again 3 encoded bits are necessary to reach the desired BER for the uncoded bits $(\tilde{m} = 2)$.

For small block lengths we will operate the coding scheme at even higher signal-to-noise ratios, so we will be on the safe side as far as this design rule is concerned. To target lower BER we will have to adjust P_b accordingly, of course, and yield a higher value for \tilde{m} .

2.3.3. Special Interleaver Design for Improved Performance. Several researchers have worked on designing good interleavers for binary Turbo codes. We would like to point out the work by Hokfelt [12-15], where the effect of the interleaver is evaluated in terms of (1) the influence on the iterative decoding algorithm and (2) the avoidance of low-weight error events. Hokfelt proposed the design of interleavers that have positive characteristics with respect to both of these criteria. Specifically, they are designed to improve the decoding performance at lower signal-to-noise ratios and simultaneously improve the BER at high signal-to-noise ratios where Turbo codes often exhibit the "flattening" of the BER curves. We have applied Hokfelt's interleavers to some of the examples

for Turbo TCM, also with marked performance gains (see Section 4).

2.4. Component Code Search

In order to find good component codes, one can perform an exhaustive computer search similar to that in [5] that maximizes the minimal distance of each component code under consideration of randomly selecting the parity bits of each second symbol. Furthermore, one should restrict the search to those codes with a primitive feedback polynomial that are widely accepted to yield good performance for Turbo codes (all codes with primitive feedback polynomial thus found have a minimal distance as good as the best candidate codes with nonprimitive feedback polynomial).

A further condition on the code is that the information bits in step k do not affect the value of the parity bits at step k; this condition was also proposed for good TCM codes [5].

Equation (15b) in [5] states that the minimal distance is bounded by

$$d_{\text{free}}^2 \ge \Delta_{\text{free}}^2 = \min \sum_{i=k}^{k+L} \Delta_{q(\mathbf{E}_i)}^2 \equiv \min \Delta^2 \left[\mathbf{E}(D) \right]$$
(2)

minimizing over all nonzero code sequences $\mathbf{E}(D)$. The variable $q(\mathbf{E}_i)$ is the number of trailing zeros in \mathbf{E}_i . The values Δ_0^2 , Δ_1^2 , Δ_2^2 , ..., are the squared minimal Euclidean distances between signals of each subset and must be replaced by $\Delta_0^{*2},~\Delta_1^{*2},~\Delta_2^{*2},~\ldots,$ when the corresponding transmitted symbol was "punctured"; the distances are shown in Fig. 3 for 8-PSK. These new distances can be calculated by assuming that the "random" parity bit takes its worst-case value and minimizes the distance between elements of the subsets. In our search we also test both possible states of the puncturing pattern (punctured, unpunctured, punctured, ..., vs. unpunctured, punctured, unpunctured, \ldots) and retain the lowest distance obtained. After such a search one obtains the results of Table 1, where the parity-check polynomials in octal notation are given as in [5]. Note that in the case of 8-PSK the punctured code has a loss compared to uncoded QPSK $(d_{\rm free}^2/d_{
m QPSK}^2=d_{\rm free}^2/2=0.878),$ but we must not forget that we are able to transmit an *additional* (parity) bit every $2 \cdot n$ 8-PSK symbols, albeit with little protection within the signal constellation.

Table 1. "Punctured" TCM Codes with Best MinimalDistance and Primitive Feedback Polynomial for 8-PSKand QAM (in Octal Notation)

Code	\tilde{m}	$H^0(D)$	$H^1(D)$	$H^2(D)$	$H^3(D)$	$d_{ m free}^2/\Delta_0^2$
2D-8-PSK, 8 states	2	11	02	04	_	3
4D-8-PSK, 8 states	2	11	06	04	_	3
2D-8-PSK, 16 states	2	23	02	10	_	3
4D-8-PSK, 16 states	2	23	14	06	_	3
$2 \mathrm{D} Z^2$, 8 states	3	11	02	04	10	2
$2D Z^2$, 16 states	3	23	02	16	04	3
$2D Z^2$, 8 states	2	11	04	02	_	3
$2D Z^2$, 16 states	2	23	04	10	_	4

3. THE DECODER

3.1. Differences to Binary Turbo Codes

The iterative decoder is similar to that used to decode binary Turbo codes, except that there is a difference in the nature of the information passed from one decoder to the other, and in the treatment of the very first decoding step. Turbo codes received their name from the iterative nature of the decoding algorithm. The main building block of the complete Turbo decoder is the component decoder, which may be a soft-output Viterbi decoder [16], or a symbol-by-symbol (S-b-S) maximum a priori (MAP) decoder [7,17,18]. The component decoders each perform optimal (or close-to-optimal) decoding of each component code, and use the output of the other component decoder as if it were an *independent* estimate of the information bits. In the binary Turbo coding scheme, it can be shown that the component decoder's output can be split into three additive parts (when in the logarithmic or loglikelihood ratio domain [17,18]) for each information bit with index k: the systematic component (corresponding to the received systematic value for bit k), the *a priori component* (the information given by the other decoder for bit k), and the extrinsic component (that part that depends on all other inputs, i.e., those to the "left" and "right" of the bit k in the associated trellis diagram and also the parity information). Only the so-called extrinsic component may be given to the next decoder; otherwise information will be used more that once in the next decoder [1,19]. Furthermore, these three components are disturbed by independent noise.

A major novelty when decoding TTCM is the fact that each decoder alternately sees its corresponding encoder's noisy output symbol(s) and then the other encoder's noisy output symbol(s). The information bits, that is, systematic bits, that partly resulted in the mapping of each of these symbols are correct-in the sense of being identical to the corresponding encoder output-in both cases. However, this is not so for all the parity bits, since these originate from the other encoder every other group of *n* symbol — we have indexed these symbols with "*," and will call these symbols "punctured" for brevity. Note that in the following, the attribute "*" or "punctured" refers to the pertinent component decoder only. The situation is further complicated by the fact that the systematic component cannot be separated from the parity one. This is because the noise that affects the parity component also affects the systematic component since (unlike in the binary case) the systematic information is transmitted together with parity information in the same symbol(s).

Fortunately, these two problems, alternating "punctured" parity bits, and inseparability of systematic and parity components can be solved simultaneously. The trick is to split the output into just two different components: (1) a priori and (2) extrinsic together with systematic. Furthermore, care is taken to avoid using the systematic information more than once in each decoder as will be explained later.

We recommend that the reader briefly review Appendix A, where we have derived the symbol-by-symbol MAP decoder for nonbinary trellises and thus to become familiar with the terms forward and backward variables (α and β), and transitions in the trellis, before continuing with Section 3.2.

3.2. Extrinsic, A Priori, and Systematic Components

Because we will now take a close look at the way the iterative decoder works, we have decided to write logarithms of probabilities, denoted by L(), for brevity and clarity. Let us thus define

- $L(d_k = i)$, the logarithm of the decoder output (A.10), written as L in diagrams.
- $L_a(d_k = i) = \log \Pr\{d_k = i\}$, the logarithm of the apriori term (A.4), written as *a* in diagrams.
- $L_{(p\&s)}(d_k = i) = \log p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M')$, the logarithm of the inseparable parity and systematic components. Note that we have written (p&s)in parentheses to stress their inseparability. It is written as (p&s) in diagrams.
- L_(e&s)(d_k = i), the logarithm of the combined extrinsic and systematic components, written as (e&s) in diagrams. It will be discussed in the following.

We had stated above that we wish to pass the combined extrinsic and systematic components, $L_{(e\&s)}$, to the next decoder in which it is used as a priori information. This part of the decoder output does not depend on the a priori information $\Pr\{d_k = i\}$. In other words, we must subtract the logarithm of the a priori term $L_a(d_k = i) = \log \Pr\{d_k = i\}$ from the logarithm of (A.10) to obtain a term independent of the a priori information $\Pr\{d_k = i\}$. Thus we compute:

$$L_{(e\&s)}(d_k = i) = \log \Pr\{d_k = i \mid \mathbf{y}\} - \log \Pr\{d_k = i\}$$
(3)

 $\forall i \in \{0, \ldots, 2^m - 1\}$. This can be done since $\Pr\{d_k = i\}$ is a factor in γ_i that does not depend on M or M' and can be written outside the summations in (A.10). Note that the parity component cannot be separated from the extrinsic once since the former depends on M and M' and cannot be written outside the summations in (A.10). Finally, since the systematic component cannot be split from the parity component, we cannot split (*e*&*s*) at all—hence the parentheses in $L_{(e\&s)}$.

However, the decoder must be formulated in such a way that it correctly uses the channel observation \mathbf{y}_k and the a priori information $Pr\{d_k = i\}$ at each step k. This is best illustrated in a diagram (see Fig. 5). Shown on the left is the interrelation of both MAP decoders for one information bit in a binary Turbo coding scheme. We have denoted the extrinsic component - omitting the index k—with e, the a priori component with a and the systematic and parity ones with s and p. Bold letters indicate that the variables correspond directly to the upper decoder, nonbold ones correspond directly to the lower decoder. Thus bold (nonbold) extrinsic and L values are produced by the upper (lower) decoder and bold (nonbold) a priori values *used* by the upper (lower) decoder. Of course, the decoders have memory (indicated by inputs α and β), so each input will affect many neighboring outputs; we have shown the relationships for only one trellis transition



Figure 5. The decoders for binary Turbo codes and TTCM. Note that the labels and arrows apply only to one specific info bit (left) or group of m info bits (right). The interleavers/deinterleavers are not shown.

(step). Both decoders are symmetric as they only pass the newly generated extrinsic information to the next decoder.

The right side of Fig. 5 shows the decoders for TTCM where the upper decoder sees a punctured symbol (which was output by the other decoder: "* mode"), in the example of the encoder in Fig. 2 it might have received a noisy observation of symbol $x_2 = 3$. The corresponding symbol from the upper encoder (2) was not transmitted. The upper decoder now ignores this symbol—indicated by the position of the upper switch—as far as the direct channel input is concerned: in Eq. (A.3) we set

$$L_{(p\&s)} = \log p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M') \to 0$$
 (4)

illustrated in Fig. 5 by $(\mathbf{p\&s}) = \mathbf{0}$. The only input for this step in the trellis is a priori information $L_{\mathbf{a}}$ from the lower decoder, which includes the systematic and (lower) parity information (p&s). The output of the MAP, for this transition, is the sum of this a-priori information $L_{\mathbf{a}}$ and newly computed extrinsic information $L_{\mathbf{e}}$, since we have set $L_{(\mathbf{p\&s})}$ to zero. The a priori information $L_{\mathbf{a}}$ is subtracted, and the extrinsic information $L_{\mathbf{e}}$ is passed to the lower decoder as *its* a-priori information, L_{a} (see the equations written in Fig. 5). The lower decoder, however, sees a symbol that *was* generated by its encoder; hence it can compute

$$L_{(p\&s)}(d_k = i) = \log p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M') \quad (5)$$

for each *i*, and subsequently $L_{(e\&s)}(d_k = i)$ using (3). Then $L_{(p\&s)}(d_k = i)$ is used as the a priori input of the upper decoder in the next iteration. The setting of the switches will alternate from one group of bits (index *k*) to another. The symmetry of the decoders seeing alternately punctured and unpunctured symbols allows decoders to include the

systematic information despite of the fact that it cannot be separated from the parity part.

3.2.1. Summary of the Decoder Iteration. We have summarized the steps of one complete iteration in the following algorithm and have numbered the steps in Fig. 6 accordingly. The horizontal line delimits the upper from the lower decoder. We begin with the left hand side of the figure which corresponds to the lower decoder seeing an unpunctured symbol:

- 1. Compute the logarithm of the branch transition probability [Eq. (A.3)] for each possible value of $d_k = i$. This now takes the systematic and parity components into account for the transition k [Eq. (5)]. Note that the last part of (A.3) denotes the a priori information ($L_a = L_e$) computed by the upper decoder associated with trellis transition k, for each possible value of $d_k = i$.
- 2. Compute the logarithms of the forward and backward variables, α_{k-1} and β_k , with Eqs. (A.1) and (A.2) for all trellis states M' and M. This now takes into account the code constraint and thus includes all a priori information generated by the upper decoder for the neighboring trellis transitions $\neq k$.
- 3. Compute the logarithm of the MAP output (A.10) for each possible value of $d_k = i$ using the results of steps 1 and 2. Then subtract the corresponding a priori information $(L_a = L_e)$ computed by the upper decoder for each possible value of $d_k = i$ associated with the transition k. The subtraction is a vector operation of length 2^m [see Eq. (3)].
- 4. Pass the result $(L_a = L_{(e\&s)})$ of step 3 to the upper decoder for it to use in its next decoding iteration. It is the combined extrinsic and systematic component.



Figure 6. Illustration of the decoder algorithm for TTCM. The symbols used are those used in the right side of Fig. 5 and explained in the text; underlined numbers refer to the steps of the detailed algorithm. The left hand refers to the case where the upper decoder sees a punctured symbol; the right hand, where the lower decoder sees a punctured symbol. The horizontal line delimits the upper decoder from the lower decoder.

In steps 5-8 the index k is the deinterleaved position of index k in steps 1-4.

- 5. Because this is a punctured symbol seen from the upper decoder's stance, compute the logarithm of the branch transition probability [Eq. (A.3)] for each possible value of $d_k = i$ setting the logarithm of $p(y_k | d_k = i, S_k = M, S_{k-1} = M')$ to zero [Eq. (4)]. Note that the last part of (A.3) denotes the a priori information $[L_a = L_{(e\&s)}]$ computed by the lower decoder associated with trellis transition k, for each possible value of $d_k = i$.
- 6. Same as step 2, exchange "upper decoder" by "lower decoder."
- 7. Compute the logarithm of the MAP output (A.10) for each possible value of $d_k = i$ using the results of steps 5 and 6. Then subtract the corresponding a priori information $[L_a = L_{(e\&s)}]$ computed by the lower decoder for each possible value of $d_k = i$ associated with the transition k. The subtraction is a vector operation of length 2^m .
- 8. Pass the result L_a of step 7 to the lower decoder to use in its next decoding iteration. Note that L_a comprises the extrinsic component L_e without any systematic or parity component.

For trellis transitions where the lower decoder sees a punctured symbol, the right side of Fig. 6 applies. The same steps (1-8) are performed, except that we swap

"upper decoder" and "lower decoder" and swap bold and nonbold notation accordingly.

For one whole iteration:

- Begin with the upper decoder.
- Use the a priori information generated by the lower decoder in its last decoding phase.
- Go through all values of k for $0 \le k < N$ applying steps 1–4 or 5–8 from above as applicable for the puncturing (seen from the upper decoder's stance) of that symbol k.
- Then go to the lower decoder.
- Use the a priori information generated by the upper decoder in its last decoding phase.
- Go through all values of k for $0 \le k < N$ applying steps 1–4 or 5–8 from above as applicable for the puncturing (seen from the lower decoder's stance) of that symbol k.

3.3. Metric Calculation in the First Decoding Stage

The description above assumes that in case a decoder sees a punctured symbol, the systematic and parity information is available from the a priori information received from the other decoder. This is the case in all except the very first decoding stage of the upper decoder. Hence, before the first decoding stage of the upper decoder, we need to set the a priori information to contain the systematic information for the * transitions, where the transmitted symbol was determined partly by the information group d_k but also by the unknown parity bit $b^{0,*} \in \{0, 1\}$ produced by the *other* encoder. We thus set the a priori information, by applying the mixed Bayes' rule, to

$$\Pr\{d_{k} = i\} \leftarrow \Pr\{d_{k} = i \mid \mathbf{y}_{k}\} = \operatorname{const} \cdot p(\mathbf{y}_{k} \mid d_{k} = i)$$
$$= \operatorname{const} \cdot \sum_{j \in \{0,1\}} p(\mathbf{y}_{k}, b_{k}^{0,*} = j \mid d_{k} = i)$$
$$= \frac{\operatorname{const}}{2} \cdot \sum_{j \in \{0,1\}} p(\mathbf{y}_{k} \mid d_{k} = i, b_{k}^{0,*} = j) \quad (6)$$

where it is assumed that $\Pr\{b_k^{0,*} = j \mid d_k\} = \Pr\{b_k^{0,*} = j\} = \frac{1}{2}$, that is, that the parity bit in the symbol x_k is statistically independent of the information bit group d_k and equally likely to be zero or one. Furthermore, the initial a priori probability of d_k —prior to any decoding—is assumed to be constant for all *i*. Above, it is not necessary to calculate the value of the constant, since the value of $\Pr\{d_k = i \mid \mathbf{y}_k\}$ can be determined by dividing the summation $\sum_{j \in \{0,1\}}$ by its

sum over all *i* (normalization). If the upper decoder is not at a * transition, then we simply set $\Pr\{d_k = i\}$ to $\frac{1}{2^m}$.

3.4. The Complete Decoder

The complete decoder is shown in Fig. 7. By 'metric s' we mean the evaluation of (6). All thin signal paths

are channel outputs or values of $\log p(\mathbf{y}_k | d_k = i, S_k = M, S_{k-1} = M')$; thick paths represent a group of 2^m values of logarithms of probabilities.

3.4.1. Avoiding Calculation of Logarithms and Exponentials. Since we work with logarithms of probabilities, it is undesirable to switch between probabilities and their logarithms. This, however, becomes necessary at the following four stages in the decoder:

1. In (6) when we sum over probabilities

$$\left(\sum_{i \in \{0,1\}} p(\mathbf{y}_k \mid d_k = i, b_k^{0,*} = j)\right), \text{ but the demodulator}$$
provides us with $\log(p(\mathbf{y}_k \mid d_k = i, b_k^{0,*} = j)).$
When evaluating $\sum \sum p(\mathbf{y}_k \mid d_k = i, b_k^{0,*} = j)$ to

normalize (6) to unity.

 $\mathbf{2}$.

- 3. When normalizing the sum of (A.10) to unity.
- 4. When calculating the hard decision of each individual bit given the values of (A.10).

All of the preceding mandate the calculation of the logarithm of the sum over exponentials (when the decoder otherwise operates in the log domain). By recursively applying the relation [18]

$$\ln(e^{\delta_1} + e^{\delta_2}) = \max(\delta_1, \delta_2) + \ln(1 + e^{-|\delta_2 - \delta_1|})$$

= max(\delta_1, \delta_2) + f_c(|\delta_1 - \delta_2|) (7)



Figure 7. The complete decoder.

the problem can be solved for an arbitrary number of exponentials. The correction function $f_c(.)$ can be realized with a one-dimensional table with as few as eight stored values [18]. When implementing the preceding, we noticed negligible performance degradation.

3.4.2. Subset Decoding. When the component code's trellis contains parallel transitions, this reduces the required decoding complexity: During the iterations, it is not necessary to decide on, or calculate soft outputs for, the uncoded bits that cause these parallel transitions. In the MAP decoders, the parallel transitions can be merged, which mathematically corresponds to adding the path transition probabilities $\gamma_i(\mathbf{y}_k, M', M)$ of the parallel transitions. It is clear that the sum is over just those $2^{(m-\tilde{m})}$ values of *i* that represent all combinations of the statistically independent uncoded bits. There is one such sum for every particular combination of the remaining \tilde{m} bits that are encoded. The MAP decoder calculates and passes on only the likelihoods of these \tilde{m} bits; hence the (de)interleaver needs to operate only on groups of \tilde{m} bits. During the very last decoding stage, decisions (and, if desired, reliabilities) for the $(m - \tilde{m})$ uncoded bits can be generated by the MAP decoder; either optimally or suboptimally, for example, by taking into account only those transitions between the most likely states along the trellis.

4. EXAMPLES AND SIMULATIONS

As examples we have used 2D-8-PSK (with N = 1024), 2D-16-QAM (with N = 683), 4D-8-PSK (with 200), and 2D-64-QAM (with N = 200 and 1024). Unless stated otherwise, the interleavers were chosen to be pseudorandom, and identical for each transmitted block. In all cases the component decoders were symbol-by-symbol MAP decoders operating in the log domain. The number of trellis states was either 4, 8, or 16. To help the reader compare curves for different values of N, the x axes of the respective curves were chosen to show the same range of SNR. The channel was modeled to be AWGN, where N_0 is the one-sided noise power spectral density. The small block sizes were included to verify that the schemes work well in applications that tolerate only short end-to-end delays. In general, it must be borne in mind that when comparing different approaches to channel coding, the block size (or other measure of fundamental delay) must be kept constant.

The BER curves resulting from Monte Carlo computer simulations are shown in Figs. 8 and 9 for 8-PSK with 2 bits per symbol (b/s); in Fig. 10 for 16-QAM with 3 b/s; in Fig. 11 for 8-PSK with 2.5 b/s and finally in Fig. 12 for 64-QAM with 5 b/s. One iteration is defined as comprising two decoding steps: one in each dimension. The weak asymptotic performance of the component code (evident from the high BER after the very first decoding step) does not seem to affect the performance of the Turbo code after a few iterations, since good BER can be achieved at less than 1 dB from Shannon's limit for large interleaver sizes N. For comparison, Fig. 8 includes the results for a Gray mapping scheme for 2D-8-PSK as presented in [3]; it has the same complexity (when measured as the number of trellis branches per information bit) as the TTCM four-iteration scheme and the same number of information bits per block: 2048. The number of states of the binary trellis for the Gray mapping scheme is 8; hence there are $2048 \times 8 \times 2$ trellis branches per decoding in each dimension; in the TTCM scheme there are $1024 \times 8 \times 4$ branches. Compared to TCM with 64-state Ungerboeck codes and 8-PSK (not included in the figures), we achieve a gain of 1.7 dB at a BER of 10^{-4} . At this BER, the TTCM system has a 0.5-dB advantage over the Gray mapping scheme after four iterations. Rather than comparing all our examples with other coding techniques, we simply point out that good BER can be achieved within 1 dB from Shannon's limit as long as the block size is sufficiently large [20]. The use of designed interleavers for 2D-8-PSK and N = 1024 is shown in Fig. 9; we see a marked improvement when using



Figure 8. TTCM for 2D-8-PSK, 2 bits per symbol (b/s). Channel capacity: 2 b/s at 5.9 dB. Random interleaver. Code with $\tilde{m} = 2$.



Figure 9. TTCM for 2D-8-PSK and different interleavers and code memory, 2 b/s. Channel capacity: 2 b/s at 5.9 dB. Code with $\tilde{m} = 2$.



Figure 10. TTCM for 2D-16-QAM, 3 b/s. Channel capacity: 3 b/s at 9.3 dB. Random interleaver. Code with $\tilde{m} = 3$.

the designed interleaver of Hokfelt [12,13], especially for 16 states.

The results for the higher-bandwidth-efficient examples (2D-64-QAM and 4D-8-PSK) are also encouraging. For most of the simulations we used a random interleaver, unadapted to the component code. This results in the characteristic flattening of the BER curves for higher signal-to-noise ratios and BER lower than 10^{-5} . This BER is consistent with our target BER for the uncoded bits when choosing \tilde{m} as explained in Section 2.3. If we target a lower BER, then we need to

1. Choose \tilde{m} such that the uncoded bits are better protected and suffer from a BER at least as good as

1000 Info bits, N = 200, 8 state code, 2.5 bit/symbol, MAP, AWGN



Figure 11. TTCM for 4D-8-PSK, 2.5 b/s. Channel capacity: 2.5 b/s at 8.8 dB. Random interleaver. Code with $\tilde{m} = 2$.

our new target BER (e.g., 10^{-7}). For 64-QAM and 4D-8-PSK this means that \tilde{m} should now be 3.

- 2. Choose component codes with 16 states (as given in Table 1).
- 3. Choose an interleaver designed for the component codes according to the technique of Hokfelt et al. [12,13].

In Fig. 13 we show results for 64-QAM modulation memory 4 component codes, 5120 information bits in 1024 blocks, using random and constructed interleavers, and after eight decoding iterations.

It is to be noted that Turbo-coded systems will often be employed as an inner coding stage by concatenating an



Figure 12. TTCM for 2D-64-QAM, 5 b/s. Channel capacity: 5 b/s at 16.2 dB. Random interleaver. Code with $\tilde{m} = 2$.

outer block code (e.g., RS or BCH code [11]) with a Turbo code, in order to reach very low BER; in these cases, BERs of around 10^{-4} are often sufficient. When we use TTCM with specially designed interleavers and achieve BERs of 10^{-7} , then only very high rate outer block codes will be needed. In the actual computer simulations performed for 2D-8-PSK and memory 4, with the designed interleaver, 2048 information bits per block, and E_S/N_0 7.5 dB, we never encountered a block with more than 11 bit errors (the majority of erroneous blocks had just 4 errors).

5. OTHER WORK AND FURTHER READING

There is a vast literature covering Turbo codes and iterative decoding. Good starting points for decoding principles and an overview of Turbo codes are [17,21], and [22]. Online resources—with further links—related to Turbo codes—can be found in [23–25].

A different approach for bandwidth-efficient coding using recursive parallel concatenation of nonbinary component codes was proposed in [26] and [27], where there is no puncturing of parity bits or symbols. A scheme using serial concatenation and four-dimensional modulation has been proposed [28]. Another serial concatenation scheme has been presented [29] that shows a marked improvement to the parallel concatenated scheme of Benedetto et al. [26,27] at high SNR but a worse performance at lower SNR, especially for larger interleavers.

Work is being carried out to assess the performance of TTCM in fading channels; see [30] and [31] for examples—the former work includes multicarrier transmission. Also, TTCM has been successfully employed in transmit antenna diversity systems in conjunction with spacetime codes in fading channels, including frequency-selective channels [32,33].

6. SUMMARY

We have illustrated the channel coding scheme called Turbo trellis-coded modulation (TTCM), which is bandwidth-efficient and allows iterative "Turbo" decoding of codes built around punctured parallel concatenated trellis codes together with higher-order modulation. The bitwise interleaver known from classic binary Turbo codes is replaced by an interleaver operating on a group of bits. By adhering to a set of constraints for component code and interleaver, the resulting code can be decoded iteratively using, for example, symbol-by-symbol MAP component decoders working in the logarithmic domain to avoid numerical problems and reduce the decoding complexity. We outlined the structure of the iterative decoder and derived the symbol-by-symbol MAP algorithm for nonbinary trellises. Furthermore, we illustrated the differences to the binary case as far as the use of extrinsic, systematic, and parity components of the symbol-bysymbol decoder output are concerned.

The search results for good component codes are shown, taking into account the puncturing at the transmitter. Simulation results are presented for codes with 4, 8, and 16 states, employing random and designed interleavers, and various signal sets such as two- and fourdimensional 8-PSK, 16-QAM, and 64-QAM. With TTCM, error correction close to Shannon's limit is possible for highly bandwidth-efficient schemes that are of relatively low complexity. It remains to be seen whether further

Figure 13. TTCM for 2D-64-QAM, 5 b/s. Channel capacity: 5 b/s at 16.2 dB. Code with $\tilde{m} = 3$.

refinement of the interleaver construction will reduce the flattening of the BER curves below 10^{-8} or 10^{-10} .

Acknowledgments

The authors would like to thank Dr. Joachim Hagenauer for valuable discussions and Dr. Johan Hokfelt for providing designed interleavers.

APPENDIX A. THE SYMBOL-BY-SYMBOL MAP ALGORITHM FOR NONBINARY TRELLISES

We will briefly show the derivation of the symbol-bysymbol MAP algorithm [7] (MAP for short) to be used as the component decoder in the iterative decoding scheme for TTCM built with for nonbinary trellises. For the derivation we consider just a conventional, unpunctured TCM scheme, with a priori information -on each group of information bits d_k - to be used at the input of the decoder. Let the number of states be 2^{ν} , and the state at step k be denoted by $S_k \in \{0, 1, \dots, 2^{\nu} - 1\}$. The group of minformation bits d_k can be represented by an integer in the range $(0 \dots 2^m - 1)$ and is associated with the transition from step k - 1 to k. The receiver observes N sets of n noisy symbols, where *n* such symbols are associated with each step in the trellis; specifically, from step k - 1 to step *k* the receiver observes $\mathbf{y}_k = [y_k^0, \dots, y_k^{(n-1)}]$. Let the total received sequence be $\mathbf{y} = \mathbf{y}_1^N = (\mathbf{y}_1, \dots, \mathbf{y}_N)$. This is the TCM encoder subset $\mathbf{y}_1 = (\mathbf{y}_1, \dots, \mathbf{y}_N)$. TCM encoder output sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ that has been disturbed by additive white Gaussian noise with one-sided noise-power spectral density N_0 . Each $\mathbf{x}_k = [x_k^0, \dots, x_k^{(n-1)}]$

is the group of *n* symbols output by the mapper at step *k*. The goal of the decoder is to evaluate $\Pr\{d_k \mid \underline{\mathbf{y}}_1^N\}$ for each d_k , and for all *k*. Let us introduce and define the so-called forward and backward variables:

$$\alpha_{k-1}(M') = \frac{p(S_{k-1} = M', \mathbf{y}_1^{k-1})}{p(\mathbf{y}_1^{k-1})}$$
(A.1)

$$\beta_k(M) = \frac{p(\mathbf{y}_{k+1}^N \mid S_k = M)}{p(\mathbf{y}_{k+1}^N \mid \mathbf{y}_1^k)}$$
(A.2)



The branch transition probability for step k, $p(d_k = i, \mathbf{y}_k, S_k = M | S_{k-1} = M')$, is denoted by, and calculated as

$$\gamma_{i}(\mathbf{y}_{k}, M', M) = p(\mathbf{y}_{k} \mid d_{k} = i, S_{k} = M, S_{k-1} = M')$$

$$\cdot q(d_{k} = i \mid S_{k} = M, S_{k-1} = M')$$

$$\cdot \Pr\{S_{k} = M \mid S_{k-1} = M'\}$$
(A.3)

 $q (d_k = i | S_k = M, S_{k-1} = M')$ is either zero or one depending on whether encoder input $i \in \{0, 1, \ldots, 2^m - 1\}$ is associated with the transition from state $S_{k-1} = M'$ to $S_k = M$. The first component of (A.3) represents the parity and systematic information available at the output of the transmission channel; its computation depends on the channel noise variance for the case of channels with additive white Gaussian noise (AWGN) [17] and on the actually transmitted symbol associated with the transition from state $S_{k-1} = M'$ to $S_k = M$ for encoder input *i*. In the last component of (A.3) we use the a priori information. For codes without any parallel transitions:

$$\begin{aligned} &\Pr\{S_{k} = M \mid S_{k-1} = M'\} = \\ &\begin{cases} &\Pr\{d_{k} = 0\} & \text{if } q(d_{k} = 0 \mid S_{k} = M, S_{k-1} = M') = 1 \\ &\Pr\{d_{k} = 1\} & \text{if } q(d_{k} = 1 \mid S_{k} = M, S_{k-1} = M') = 1 \\ &\vdots & \cdots \\ &\Pr\{d_{k} = 2^{m} - 1\} & \text{if } q(d_{k} = 2^{m} - 1 \mid S_{k} = M, \\ & S_{k-1} = M') = 1 \end{aligned}$$

$$&= \Pr\{d_{k} = j\}, \end{aligned}$$
(A.4)

where $j:q(d_k = j | S_k = M, S_{k-1} = M') = 1$. Naturally, when $q(d_k = i | S_k = M, S_{k-1} = M') = 1$, then j = i; otherwise the value of $\Pr\{S_k = M | S_{k-1} = M'\}$ and hence j will be irrelevant anyway. Formally, if there does not exist a j such that $q(d_k = j | S_k = M, S_{k-1} = M') = 1$, then $\Pr\{S_k = M | S_{k-1} = M'\}$ is set to zero.

We shall now try to combine (A.1), (A.2), and (A.4). We must first bear in mind that the event $(d_k = i, \mathbf{y}_k, S_{k-1} =$

M') has no influence on $\underline{\mathbf{y}_{k+1}^N}$ if S_k is given; hence we can write

$$p(\underline{\mathbf{y}_{k+1}^{N}} \mid S_k = M) = p(\underline{\mathbf{y}_{k+1}^{N}} \mid d_k = i, \mathbf{y}_k, S_k = M, S_{k-1} = M')$$
(A.5)

Using (A.5) and the fact that

$$p(\underline{\mathbf{y}_{1}^{k-1}}) = \frac{p(\underline{\mathbf{y}_{1}^{k}})}{p(\mathbf{y}_{k} \mid \underline{\mathbf{y}_{1}^{k-1}})}$$
(A.6)

the product of (A.1), (A.2), and (A.4) can now be shown to be:

$$\begin{aligned} \alpha_{k-1}(M') & \cdot \beta_k(M) \cdot \gamma_i(\mathbf{y}_k, M', M) \\ &= p(S_{k-1} = M', \underline{\mathbf{y}_1^{k-1}}) \\ & \cdot p(\underline{\mathbf{y}_{k+1}^N}, d_k = i, \mathbf{y}_k, S_k = M \mid S_{k-1} = M') \\ & \cdot \frac{p(\mathbf{y}_k \mid \underline{\mathbf{y}_1^{k-1}})}{p(\mathbf{y}_1^N)} \end{aligned}$$
(A.7)

Obviously

$$p(\underline{\mathbf{y}_{k+1}^{N}}, d_{k} = i, \mathbf{y}_{k}, S_{k} = M \mid S_{k-1} = M')$$

= $p(\underline{\mathbf{y}_{k+1}^{N}}, d_{k} = i, \mathbf{y}_{k}, S_{k} = M \mid S_{k-1} = M', \underline{\mathbf{y}_{1}^{k-1}})$ (A.8)

so we can re-write (A.7) as

$$\begin{aligned} \alpha_{k-1}(M') &\cdot \beta_k(M) \cdot \gamma_i(\mathbf{y}_k, M', M) \cdot \frac{1}{p(\mathbf{y}_k \mid \underline{\mathbf{y}_1^{k-1}})} \\ &= p(S_{k-1} = M', S_k = M, d_k = i, \underline{\mathbf{y}_1^N}) \frac{1}{p(\underline{\mathbf{y}_1^N})} \\ &= p(S_{k-1} = M', S_k = M, d_k = i \mid \underline{\mathbf{y}_1^N}) \end{aligned}$$
(A.9)

Therefore, the desired output of the MAP decoder is

$$\begin{aligned} \Pr\{d_k = i \mid \underline{\mathbf{y}}\} &= \text{const} \ \cdot \sum_{M} \sum_{M'} \gamma_i(\mathbf{y}_k, M', M) \\ &\cdot \alpha_{k-1}(M') \cdot \beta_k(M) \end{aligned} \tag{A.10}$$

 $\forall i \in \{0, \ldots, 2^m - 1\}$. The constant can be eliminated by normalizing the sum of (A.10) over all *i* to unity. The probability $\Pr\{d_k = i \mid \underline{y}\}$ comprises a priori, systematic, parity, and extrinsic components, since it depends on the complete received sequence as well as the a priori likelihoods of d_k .

All that remains now is to recursively define $\alpha_{k-1}(M')$ and $\beta_k(M)$. We begin by writing

$$\Pr\{S_k = M \mid \underline{\mathbf{y}_1^{k-1}}, \mathbf{y}_k\} \cdot p(\mathbf{y}_k \mid \underline{\mathbf{y}_1^{k-1}})$$
$$= p(\mathbf{y}_k, S_k = M \mid \underline{\mathbf{y}_1^{k-1}})$$
(A.11)

and dividing both sides by $p(\mathbf{y}_k \mid \underline{\mathbf{y}_1^{k-1}})$ and expanding into the form

$$\Pr\{S_k = M \mid \mathbf{y}_1^k\} = \alpha_k(M)$$

$$= \frac{\sum_{M'} p(\mathbf{y}_k, S_k = M, S_{k-1} = M' \mid \underline{\mathbf{y}_1^{k-1}})}{\sum_{M} \sum_{M'} p(S_k = M, S_{k-1} = M', \mathbf{y}_k \mid \underline{\mathbf{y}_1^{k-1}})}$$
(A.12)

Because of (A.8), we can write

$$\alpha_{k}(M) = \frac{\sum_{M'} Pr\{\mathbf{y}_{k}, S_{k} = M \mid S_{k-1} = M'\}}{\sum_{M} \sum_{M'} p(\mathbf{y}_{k}, S_{k} = M \mid \mathbf{y}_{1}^{k-1})} \quad (A.13)$$
$$\cdot \Pr\{S_{k-1} = M' \mid \mathbf{y}_{1}^{k-1}\}$$

Defining

$$\gamma_T(\mathbf{y}_k, M', M) = \sum_{i=0}^{2^m - 1} \gamma_i(\mathbf{y}_k, M', M)$$
(A.14)

yields

$$\alpha_{k}(M) = \frac{\sum_{M'} \gamma_{T}(\mathbf{y}_{k}, M', M) \cdot \alpha_{k-1}(M')}{\sum_{M} \sum_{M'} \gamma_{T}(\mathbf{y}_{k}, M', M) \cdot \alpha_{k-1}(M')}$$
(A.15)

Similarly

$$\beta_{k}(M) = \frac{\sum_{M''} p(S_{k+1} = M'', \underline{\mathbf{y}_{k+1}^{N}} | S_{k} = M)}{p(\underline{\mathbf{y}_{k+1}^{N}} | \underline{\mathbf{y}_{1}^{k}})}$$
$$= \frac{\sum_{M''} p(S_{k+1} = M'', \mathbf{y}_{k+1} | S_{k} = M)}{p(\underline{\mathbf{y}_{k+2}^{N}} | S_{k+1} = M'')}$$
$$= \frac{p(\underline{\mathbf{y}_{k+2}^{N}} | \underline{\mathbf{y}_{1}^{k}})}{p(\underline{\mathbf{y}_{k+1}^{N}} | \underline{\mathbf{y}_{1}^{k}})} \quad (A.16)$$

since $p(\mathbf{y}_{k+2}^N \mid S_{k+1} = M'') = p(\mathbf{y}_{k+2}^N \mid S_{k+1} = M'', \mathbf{y}_{k+1}, S_k = M)$. Finally, we can calculate $\beta_k(M)$ recursively using

 $\beta_k(M)$

$$= \frac{\sum_{M''} p(S_{k+1} = M'', \mathbf{y}_{k+1} | S_k = M) \cdot \frac{p(\underline{\mathbf{y}_{k+2}^N} | S_{k+1} = M'')}{p(\underline{\mathbf{y}_{k+2}^N} | \underline{\mathbf{y}_1^{k+1}})}}{\sum_{M''} \sum_{M} p(S_{k+1} = M'', S_k = M, \mathbf{y}_{k+1} | \underline{\mathbf{y}_1^k})}$$
$$= \frac{\sum_{M''} \gamma_T(\mathbf{y}_{k+1}, M, M'') \cdot \beta_{k+1}(M'')}{\sum_{M''} \sum_{M} \gamma_T(\mathbf{y}_{k+1}, M, M'') \cdot \alpha_k(M)}$$
(A.17)

In our implementation of the preceding algorithm, we have used logarithms of probabilities and logarithms of $\alpha_{k-1}(M')$, $\beta_k(M)$, and $\gamma_i(\mathbf{y}_k, M', M)$ employing the quasioptimal log-MAP algorithm [18] that uses the max function in conjunction with a table lookup to compute the logarithm of a sum of exponentials. The loss incurred through the use of the log-MAP algorithm is less than 0.1 dB even when using a lookup table with eight stored values.

2752 TURBO TRELLIS-CODED MODULATION (TTCM) EMPLOYING PARITY BIT PUNCTURING AND PARALLEL CONCATENATION

BIOGRAPHIES

Patrick Robertson received the Dipl.-Ing. degree in electrical engineering from the Technical University of Munich, Munich, Germany, in 1989 and a Ph.D. from the University of the Federal Armed Forces, Munich, Germany, in 1995. Since 1990, he has been working at the Institute for Communications Technology at the German Aerospace Centre (DLR). From 1990 to 1993 he shared this position with a part-time teaching post at the University of the Federal Armed Forces. His contributions within a number of EU and national R&D projects have included work on key technical definition and standardization, such as for the DVB-T system for digital terrestrial television transmission. In 1999 he became leader of the research Group "Broadband Systems and Navigation." Dr. Robertson's current interests include wireless communications and channel coding, navigation systems, navigation channel measurement, as well as mobile multimedia platforms and service architectures. He holds several patents and has published numerous research papers on these areas.

Thomas Worz was born in Stuttgart, Germany, in 1961. He received the Dipl. Ing. degree in electrical engineering from the Technical University of Stuttgart, Germany, in 1988 and his Ph.D. from the Technical University of Munich, Munich, Germany, in 1995. Since 1988, he has been with the Institute of Communications Technology of the German Aerospace Center (DLR), Oberpfaffenhofen. In 1991, he spent a three-month period as a guest scientist at the Communications Research Centre (CRC), Ottawa, Canada. In 1999, he cofounded the AUDENS Advanced Communications Technology Consulting GmbH and works as a technical consultant to industry and agencies. His research interests include channel coding, coded modulation, synchronization, signal processing, and system design. Currently, he is involved in the definition of the Galileo European Navigation system.

BIBLIOGRAPHY

- C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. ICC'93*, May 1993, pp. 1064–1070.
- 3GPP, Universal Mobile Telecommunications System (UMTS) Multiplexing and Channel Coding (FDD), No. 125.212, v.4.1.0 (standard), 2001.
- S. Le Goff, A. Glavieux, and C. Berrou, Turbo-codes and high spectral efficiency modulation, *Proc. ICC'94*, May 1994, pp. 645-649.
- U. Wachsmann and J. Huber, Power and bandwidth efficient digital communication using turbo codes in multilevel codes, *Eur. Trans. Telecommun.* 6(5): (1995).
- G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**: 55–67 (Jan. 1982).
- P. Robertson and T. Woerz, Bandwidth efficient turbo trellis coded modulation using punctured component codes, *IEEE J. Select. Areas Telecommun.* 16: (Feb. 1998).

- L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).
- S. Pietrobon et al., Trellis-coded multidimensional phase modulation, *IEEE Trans. Inform. Theory* **IT-36**: 63-89 (Jan. 1990).
- S. Benedetto and G. Montorsi, Performance evaluation of parallel concatenated codes, *Proc. ICC'95*, June 1995, pp. 663-667.
- W. Blackert and S. Wilson, Turbo trellis code modulation, Proc. CISS'96, 1996.
- J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
- J. Hokfelt, O. Edfors, and T. Maseng, Interleaver design for turbo codes based on the performance of iterative decoding, *Proc. ICC'99*, June 1999.
- J. Hokfelt, O. Edfors, and T. Maseng, On the theory and performance of trellis termination methods for turbo codes, *IEEE J. Select. Areas Commun.* 19: (May 2001).
- J. Hokfelt, On the Design of Turbo Codes, Ph.D. thesis, Lund Univ., Sweden, Aug. 2000 (published by Dept. Advanced Electronics, ISBN 91-7874-061-4).
- J. Hokfelt, J. Hokfelt's Turbo codes Internet homepage (online), http://www.tde.lth.se/home/jht/index.html (Oct. 2001).
- J. Hagenauer and P. Hoeher, A Viterbi algorithm with softdecision outputs and its applications, *Proc. GLOBECOM'89*, Nov. 1989, pp. 1680–1686.
- J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. IT* (March 1996).
- P. Robertson, P. Hoeher, and E. Villebrun, Optimal and suboptimal maximum a posteriori algorithms suitable for turbo decoding, *Eur. Trans. Telecommun.* 8(2): 119-125 (1997).
- J. H. Lodge, R. Young, P. Hoeher, and J. Hagenauer, Separable MAP 'filters' for the decoding of product and concatenated codes, *Proc. ICC'93*, May 1993, pp. 1740–1745.
- P. Robertson, An overview of bandwidth efficient turbo coding schemes, Proc. Int. Symp. Turbo Codes and Related Topics, Sept. 1997, pp. 103-110.
- S. ten Brink, Convergence behaviour of iteratively decoded parallel concatenated codes, *IEEE Trans. Commun.* 49: 1727-1737 (Oct. 2001).
- J. Woodard and L. Hanzo, Comparative study of turbo decoding techniques: An overview, *IEEE Trans. Vehic. Technol.* 49: 2208-2233 (Nov. 2000).
- 23. S. Pietrobon, ITR's Turbo coding home page (online), http://www.itr.unisa.edu.au/steven/turbo/ (Oct. 2001).
- R. Pyndiah, Block turbo codes (online), http://www-sc.enstbretagne.fr/turbo/principale.html (Oct. 2001).
- M. Valenti, Turbo codes at West Virginia University (online), http://www.csee.wvu.edu/mvalenti/turbo.html (Oct. 2001).
- S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Bandwidth efficient parallel concatenated coding schemes, *Electron. Lett.* 31(24): 2067–2069 (1995).
- S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Parallel concatenated trellis coded modulation, *Proc. ICC'96*, June 1996, pp. 974–978.

TURBO TRELLIS-CODED MODULATION (TTCM) EMPLOYING PARITY BIT PUNCTURING AND PARALLEL CONCATENATION 2753

- D. Divsalar and F. Pollara, Serial and hybrid concatenated codes with applications, Proc. Int. Symp. Turbo Codes and Related Topics, Sept. 1997, pp. 80–87.
- 29. H. Ogiwara and V. Bajo, Iterative decoding of serially concatenated punctured trellis-coded modulation, *IEICE Trans. Fund.* (special section on information theory and its applications) **E82-A**: 2089–2095 (Oct. 1999).
- L. Piazzo and L. Hanzo, TTCM-OFDM over dispersive fading channels, *Proc. VTC*'2000, May 2000.
- N. Ha and R. Rajatheva, Performance of turbo trellis coded modulation (T-TCM) on frequency-selective Rayleigh fading channels, *Proc. ICC'99*, June 1999.
- 32. G. Bauch, J. Hagenauer, and N. Seshadri, Turbo processing in transmit antenna diversity systems, *Ann. Telecommun.* (Special Issue: Turbo Codes—a Widespreading Technique)
 56: 455–471 (Aug. 2001).
- G. Bauch, Concatenation of space-time block codes and Turbo-TCM, Proc. ICC'99, June 1999, pp. 1202–1206.

U

ULTRAWIDEBAND RADIO

KAZIMIERZ (KAI) SIWIAK LAURA L. HUCKABEE Time Domain Corporation Huntsville, Alabama

1. INTRODUCTION

Ultrawideband signaling is essentially the art of generating, modulating, emitting, and detecting baseband digital signals that inherently occupy large bandwidths. Impulse transmissions date back to the infancy of wireless technology. They include the experiments of Heinrich Hertz in the 1880s, and the 100-year-old spark-gap "impulse" transmissions of Guglielmo Marconi, who in 1901 sent the first ever over-the-horizon wireless transmission from the Isle of Wight to Cornwall on the British mainland. Early radio circuits consisted solely of passive electrical components, no tubes or transistors, and hence lacked the means to efficiently deal with short transient impulses. Therefore radio subsequently developed along narrowband frequencyselective analog techniques. This led to voice broadcasting and telephony—and more recently to digital telephony and wireless data. Through the years, a small cadre of scientists have worked to develop and refine impulse technologies. Before 1970 the primary focus in impulse radio research was on impulse radar techniques and government-sponsored projects. In late 1970s and early 1980s, however, digital techniques began to mature to the point where the practicality of modern low-power impulse radiocommunications could be demonstrated using the impulse time coding and time modulation approach. Digital impulse radio [1–9], the modern echo of Marconi's century-old transmissions, now emerges under the banner "ultrawideband" radio. Alternate methods of generating signals having UWB characteristics are being developed including the use of continuous streams of pseudonoise (PN)-coded impulses that resemble code-division multiple access (CDMA) signaling that employ a chip rate commensurate with the emission center frequency. The industry is now moving to commercial deployment.

UWB signaling is more nearly characterized by transient circuit responses, whereas conventional radio tends to deal with the steady state. Impulse propagation, especially indoors, also differs significantly from that of narrowband carrier-based systems in that multipath is described as distinct short impulses that sometimes overlap rather than continuous sine waves that form complex interference patterns. Many applications of UWB systems have been enabled by the unique characteristics of this technology.

2. CODED UWB IMPULSES AND IMPULSE STREAMS

UWB radio is the transmission and reception of ultrashort electromagnetic energy impulses. It is the generic term describing radio systems having very large instantaneous bandwidths. The U.S. Federal Communications Commission (FCC), for example, has tentatively defined UWB systems as "having bandwidths greater than 25% of the center frequency measured at the 10 dB down points" or "RF bandwidths greater than 1.5 GHz," whichever is smaller. There are several methods of generating, radiating and receiving such UWB signals, including TM-UWB, DS-UWB, and TRD-UWB. Wide spectra are generated in each method, however, radio techniques, signal characteristics, and application capabilities vary considerably.

Developers of UWB technology have perfected various ways for creating and receiving these signals, and for encoding information in the transmissions. Pulses can be sent individually, in bursts, or in near-continuous streams, and they can encode information in pulse amplitude, polarity, and position. Modulations vary from simple pulse position, to a more energy-efficient pulse polarity [10], and to the very-energy-efficient M-ary (multilevel) pulse position modulation. Modern UWB radio is characterized by very low effective radiated power (in the submilliwatt range) and extremely low power spectral densities, by virtue of the wide bandwidths (>1 GHz). The emissions are targeted to be below an effective isotropicaly radiated power (EIRP) of -41.25 dBm/MHz, with restrictions, in bands below 960 MHz, between 1.99 GHz and 10.6 GHz, and at 24 GHz, under U.S. CFR-47 Part 15 Report and Order issued February 2002.

The following three commercially useful UWB communications techniques exemplify the wide range of implementation possibilities: TM-UWB, DS-UWB, and TRD-UWB. All systems use transient switching techniques to generate brief (typically subnanosecond) impulses or "monocycles" having a small number of zero crossings. The impulses are radiated by specialized wideband antennas [11].

TM-UWB impulses are transmitted at high rates, in the millions to tens of millions of impulses per second. However, the pulses are not necessarily evenly spaced in time, but rather they may be spaced at random or pseudorandom time intervals. The process creates a noiselike signal in both the time and frequency domains. Data modulation is applied by further dithering the timing of the pulse transmissions, by signal polarity and perhaps pulse amplitude. A coherent correlation-type receiver and integrator converts the UWB pulses to a baseband digital signal that has a bandwidth commensurate with the data rate. The correlation operation and subsequent integration filtering provide significant processing gain, which is effective against interference and jamming. Time coding of the pulses allows for channelization, while the time dithering, pulse position, and signal polarity provide the modulation. UWB systems built around this technique and operating at very low RF power levels have demonstrated very impressive short- and long-range data links, positioning measurements accurate to within a few

centimeters, and high-performance through-wall motion sensing radars.

DS-UWB uses high duty-cycle phase-coded sequences of wideband impulses transmitted at gigahertz rates. Sequences of tens to thousands of impulse "chips" encode data bits in scalable data rates from a one to hundreds of Mbps (megabits per second). The modulation is by pulse polarity and resembles a baseband binary phaseshift-keyed (BPSK) CDMA system with the chipping rate commensurate with the center frequency. The PN (pseudonoise) encoding per data bit provides a measure of multipath delay spread tolerance, allows for channelization, and provides processing gain against interferers. A direct sequence-type of receiver can be used to correlate with the PN code and convert the integrated impulses to data rate bandwidths.

TRD-UWB employs impulse pairs that are differentially polarity encoded by the data with the transmitted pulse pairs having a precise spacing D. The receiver comprises a correlator with one input fed directly and another input delayed by D. It is similar to a conventional differential phase-shift-keyed (DPSK) system, except that rather than integrating over a bit time, here the integration time is commensurate with multipath decay time. The differentially encoded delayed-reference impulse and its data impulse are affected in the same way by multipath. Hence, the delayed-reference detection and integration operation can be made to behave like a near-perfect RAKE receiver capturing a large percentage of the multipath-induced signal echoes.

3. UWB TECHNOLOGY BASICS

UWB spectra can be generated in several different ways, such as by TM-UWB, which uses low duty-cycle impulses; by DS-UWB, which uses high-duty-cycle waveforms that are direct-sequence phase-modulated; and by coded pulse pairs in TRD-UWB. Ultra-short-impulse waveforms are common to both technologies.

The monocycle waveform applied to the transmitting antenna, and represented in Fig. 1 along with its frequency spectrum, is the most basic element of UWB signaling. A useful analytic representation of the monopulse wave form is given by p(t), the first time derivative of a Gaussian monocycle pulse

$$p(t) = -2\pi f_c t \exp\left[\frac{1}{2}\{1 - (2\pi f_c t)^2\}\right]$$
(1)

where the center frequency is f_c . The spectrum of p(t) is given by

$$P(f) = \frac{f}{f_c} \exp\left[\frac{1}{2}\left[1 - \left(\frac{f}{f_c}\right)^2\right]\right]$$
(2)

Actual radiated waveforms and spectra, like the E field shown in Fig. 1, as well as the received waveform and spectra, are further shaped by the bandpass and transient response characteristics of the transmitting antenna. When the transmitting antenna has a wideband and linear phase response, the radiated waveform approximately resembles the time derivative of the signal supplied to the transmitting antenna. The waveform and its spectrum change again in the receiving antenna load, reflecting the transient impulse response of the entire UWB radio link.

If the pulses had been sent at a regular interval without PN encoding, the resulting spectrum would contain "comb lines" separated by the inverse of the pulse repetition rate. The resulting peak power in the comb lines would limit the total transmit power undesirably, as measured in any 1-MHz bandwidth. To make the spectrum more noiselike and provide for channelization in TM-UWB, the monocycle impulses are pseudorandomly placed within each timeframe.

TM-UWB employs PN encoded time dithering to place pulses to picosecond accuracy within a time window equal to the inverse of the average pulse repetition rate.



Figure 1. Source, emitted, and received monopulses. (After Ref. 12.)



Figure 2. PN-coded UWB waveform sequence in time.



Figure 3. PN coded UWB waveform sequence in frequency.

Figure 2 illustrates a "pulsetrain" that has been PN-timecoded, and Fig. 3 shows the resulting noiselike frequency spectrum. The PN coding uses a pseudorandom timeshift within each time frame. DS-UWB, on the other hand, uses PN codes to polarity-modulate pulse sequences that are closely spaced and at regular intervals. TRD-UWB uses precisely spaced pulse pairs that are polarity-modulated. The resulting spectra of DS-UWB and TRD-UWB are similar to those of TM-UWB.

3.1. TM-UWB Technology

TM-UWB transmitters emit ultrashort monocycle waveforms with tightly controlled pulse-to-pulse intervals. The waveform pulsewidths are typically between 0.2 and 1 ns, corresponding to center frequencies between 5 and 1 GHz, with pulse-to-pulse intervals of 25-1000 ns. The systems typically use pulse position and polarity modulation. The pulse-to-pulse interval is varied on a pulse-by-pulse basis in accordance with two components: an information signal and a channel code. The TM-UWB receiver directly converts the received RF signal into a baseband digital or analog output signal. A front-end correlator coherently converts the electromagnetic pulsetrain to a baseband signal in one stage. There is no intermediate-frequency stage, greatly reducing complexity. A single bit of information may spread over multiple monocycles, providing a way of scaling the energy content of a data bit with the data rate. The receiver coherently sums the proper number of pulses to recover the transmitted information.

TM-UWB systems use a fine pulseshift modulation by positioning the pulse one quarter-cycle (60 ps for a 240-ps pulse) early or late relative to the nominal PN-coded location, or by pulse polarity. Furthermore, multilevel pulse position modulation may be used to provide enhanced bit-energy-to-noise ratio performance. The error probability P_F of the fine-shift modulation in additive white Gaussian noise (AWGN) follows the same behavior as conventional orthogonal or on/off keying (OOK)

$$P_F = \frac{1}{2} \operatorname{erfc}\left(\sqrt{\frac{\gamma_b}{2}}\right) \tag{3}$$

where γ_b is the received signal-to-noise ratio (SNR) per information bit. The error probability P_P of pulse polarity modulation in AWGN follows the same behavior as that of conventional BPSK or antipodal signaling

$$P_p = \frac{1}{2} \operatorname{erfc}(\sqrt{\gamma_b}) \tag{4}$$

where γ_b is the SNR per information bit. Pulses may also be transmitted in a "one of many positions" *M*-ary pulse position modulation, which, if the impulse positions do not overlap, resembles the performance of conventional *M*-ary orthogonal signaling in AWGN. The probability of a symbol error is

$$P_{m} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[1 - \left(1 - \frac{1}{2} \operatorname{erfc}\left(\frac{y}{\sqrt{2}}\right) \right)^{M-1} \right] \\ \exp\left[\frac{-(y - \sqrt{2\gamma})^{2}}{2} \right] dy$$
(5)

where $\gamma = \gamma_b \log(M) / \log(2)$ and γ_b is the received SNR per information bit, and average bit error probability P_M is then

$$P_M = P_m \frac{M}{2(M-1)} \tag{6}$$

Modulation further "smooths" the signal spectrum, thus making the signals more noiselike. The probability of a bit error as a function of SNR per bit for the various modulations is depicted in Fig. 4. See also Ref. 13 for derivations of P_F , P_P , and P_M .



Figure 4. Probability of error for various UWB modulations.



Figure 5. A TM-UWB transmitter.

3.2. A TM-UWB Transmitter

Figure 5 shows a high-level block diagram of a TM-UWB transmitter. The transmitter has no power amplifier, but rather, pulses are generated at the required power. A precision-programmable delay implements the PN time coding and both fine and *M*-ary pulse/time position modulation. Alternatively or in addition, modulation can be encoded in pulse polarity. The precise timing capability of the timer operation (several picoseconds resolution) enables not only precise time modulation and precise PN encoding, but also precision distance determination. The picosecond precision timer, implemented in an integrated circuit, is a key technological component of the TM-UWB system.

3.3. A TM-UWB Receiver

The receiver shown in Fig. 6 resembles the transmitter, except that the pulse generator feeds the multiplier within the correlator. The performance of this type of correlator receiver is described in Ref. 12. Baseband signal processing extracts the modulation and controls signal acquisition and tracking. Baseband signal processing also



Figure 6. A TM-UWB receiver.

drives a tracking loop that locks onto the time-coded sequence. Modulation is decoded as either an "early" or "late" pulse in time modulation and/or as a positive or negative pulse in polarity modulation. Different PN time codes are used for channelization. Precise pulse timing inherently enables exceptional positioning and location capabilities in TM-UWB communications systems.

3.4. DS-UWB Technology

A second method of generating useful signals having UWB spectra represents a DS-UWB approach, similar to an RF carrier-based CDMA system. Impulse sequences at duty cycles approaching that of a sine-wave carrier are direct-sequence polarity-modulated (like binary-phaseshift-keying). The PN sequence provides smoothing, channelization and modulation. The chipping rate is some fraction 1/N (N need not be an integer) of the "carrier" center frequency. For illustration, Fig. 7 shows the approximate spectral envelope of a 4 GHz impulse sequence that is DS modulated by a zero mean PN code for the cases N = 1 and N = 2. Actual PN sequences are relatively short and the spectra contain more features, as depicted in Fig. 3. Both signals in Fig. 7 have the same power in a 1-MHz bandwidth at 4 GHz, but the N = 1signal carries the greater total power in the spectrum. The total power and occupied bandwidth can be traded off subject to regulatory emissions limits.

3.5. TRD-UWB Technology

A method of transmitting and receiving impulses that can implement a near-perfect RAKE receiver is exemplified by TRD-UWB and described in Ref. 14. The method employs differentially encoded impulse pairs sent at a precise spacing *D*. The system is shown in the simplified block diagram of Fig. 8. The transmitter sends a pair of pulses separated by a delay *D*, and differentially encoded by pulse polarity. The pulses, including propagation induced multipath replicas, are received and detected using a correlator with one input fed directly and another input delayed by *D*. The receiver resembles a conventional DPSK receiver, which in AWGN exhibits an error probability P_D [13] of

$$P_D = \frac{1}{2} \exp\left(-\gamma_b \frac{N-1}{N}\right) \tag{7}$$



Figure 7. Spectral envelope of DS-UWB signals with N = 1 and N = 2.

where N > 1 is the number of differentially encoded pulses in a sequence and γ_b is the SNR per bit. The integration interval is sufficiently long to RAKE in a significant amount of the multipath energy. The TRD-UWB receiver tends to behave like a near-perfect RAKE receiver capturing a large percentage of the multipath induced signal echoes.

One channelization method employing TRD-UWB [14] has N = 2 and employs a family of delays D_i . Impulse pair sequences of these delay combinations constitute the channels. Figure 4 compares the error probability P_D of TRD-UWB with N arbitrarily large to the performance of other modulations.

4. UWB SIGNAL PROPAGATION

Some UWB techniques thrive in multipath, enabling positioning accuracies to better than a few centimeters, and generally follow a free-space propagation law [15]. Further indoor channel characteristics are described elsewhere [16,17]. Multipath fading, characteristic in conventional RF communications, is the result of coherent interaction of sinusoidal signals arriving by many paths. Spread spectrum TIA/EIA-95-B cellular and PCS systems with a 1.228-MHz spreading bandwidth can resolve multipath signals having differential delays of slightly less than one microsecond. Some communications channels, particularly outdoors, can have rms delay spreads measuring many microseconds; therefore, some multipath components can be resolved and received using RAKE techniques. However, in-building communications channels exhibit multipath differential delays and rms delay spreads in the several tens of nanoseconds as seen in Fig. 9 and cannot be resolved in the relatively narrow TIA/EIA-95-B channel. Those systems must therefore contend with significant Rayleigh fading, which may



Figure 8. A TRD-UWB transmitter (a) and receiver (b).



Figure 9. Measured RMS delay spread in small offices and homes.

require signals up to tens of decibels above the static signal level for a given measure of performance.

4.1. In-Building Propagation of Impulses

UWB signal propagation in free space between two unity gain antennas separated by d is very nearly

$$P_L = 20 \log\left(\frac{c}{4\pi df_c}\right) \tag{8}$$

where *c* is the velocity of light and f_c is the center frequency of the emitted spectrum. The frequency dependence of propagation comes from the frequency dependence of a unity-gain receive antenna aperture area $A_e = c^2/(4\pi f_c^2)$ evaluated here at f_c . Equation (8) is only approximately correct for the large bandwidth UWB signal [12], because total received power involves an integral in frequency over the product of the received power spectral density and A_e . A typical UWB impulse subjected to multipath is shown in Fig. 10. The multipath is evident as delayed echoes of the first-arriving impulse adding to the signal voltage. This measurement, and subsequent propagation measurements were gathered using Time Domain Corporation's pulson application demonstrator (PAD) radios, which have a built-in waveform scanning mode capable of resolving impulses to a fraction of a nanosecond.

Signal measurements in several multipath office and home environments using PADs were processed to determine the strongest impulse in a waveform versus distance and are shown in Fig. 11. The signal attenuation is shown relative to the d = 1 meter signal level. The median attenuation with distance is approximately 29 $\log(d)$ with the distance d in meters, and is typical of what can also be expected for conventional narrowband channels.

The measurements reprocessed to "perfectly RAKE" the total power in all of the multipath impulses waveform echoes, are shown in Fig. 12. The total power median attenuation with distance follows a square law, $20 \log(d)$. A "perfect RAKE" receiver can provide up to the limit of the difference between the strongest impulse and total





Figure 11. Strongest received UWB impulse versus distance. (After Ref. 12.)



Figure 12. Total UWB impulse power versus distance. (After Ref. 12.)

power, or approximately an average $9\log(d)$ of RAKE gain for the measured environments considered.

With perfect RAKE gain the signal in multipath follows a near-free-space propagation law, and represents one of the special benefits of UWB impulse technology.

Figure 10. Typical measured UWB received signal in low multipath.

4.2. Impulse Propagation with a Ground Reflection

Propagation over a smooth earth involves a reflection from the ground (Fig. 13). The direct path and reflected pathlengths D and R in terms of the antenna heights H_1 and H_2 , and the separation distance d are

$$D = \sqrt{d^2 + (H_1 - H_2)^2} \tag{9}$$

and

$$R = \sqrt{d^2 + (H_1 + H_2)^2} \tag{10}$$

See Ref. 18 for details. The differential delay between the reflected path and the direct path over a plane earth is

$$\Delta t = \frac{R - D}{c} \tag{11}$$

where *c* is the velocity of light.

The ground reflection coefficient for the cases of interest (shallow incidence angles) is very nearly -1, so the reflected pulse undergoes a polarity inversion. Reflected pulses that arrive by paths having differential delays greater than a half pulselength, as portrayed in Fig. 14, add to the total received energy. Pulses with a differential



Figure 13. Geometry for two-path propagation.



Figure 14. Overlapping and nonoverlapping pulses.

delay of *less* than half a pulselength begin to exhibit destructive interference in the receive window.

The overlapping pulse echo arriving by ground reflection is not delayed enough to be distinct from the directly arriving pulse. The nonoverlapping pulse is distinct, and adds to the received energy if a RAKE receiver is employed. With RAKE gain, the bold line in Fig. 15 in the "no-overlap region" would be raised by 3 dB. Overlapping pulses, as seen in Fig. 15, at first add constructively when the overlap is less than a half pulselength, then when nearly fully overlapping exhibit an inverse 4th power with distance behavior similar to harmonic wave propagation. In contrast, harmonic waves exhibit multiple constructive and destructive interferences as multiple sinusoidal cycle delays interact at close distances.

4.3. Reception of UWB Impulses

UWB signals are detected in a correlation-type receiver, (Figs. 6 and 8). A filter with impulse response h(t) is optionally placed between signal s(t) at the receiver antenna load and the correlator input. The correlation template pulse p(t), locally generated in Fig. 6 and derived from the transmitted reference pulse in Fig. 8, multiplies the received data pulse and is integrated and sampled at the correlator output. The receiver implementation efficiency e_c [12] of this operation is

$$e_{c} = 10 \log \left[\frac{\left| \int \int s(\tau) h(\tau - t) d\tau p(t) dt \right|^{2}}{\int s(t)^{2} dt \int \left| \int p(\tau) h(\tau - t) d\tau \right|^{2} dt} \right]$$
(12)

and is maximized when

$$C\int p(\tau)h(\tau-t)d\tau = s(t)$$
(13)

provided h(t) is causal and where *C* is the rms value of s(t). Solutions to Eq. (13) range from the matched template, $h(t) = \delta(t)$ [the Dirac delta function] with p(t) = s(t), to



Figure 15. Impulse (bold) and harmonic wave propagation near ground.

the matched filter, h(t) = s(-t) with $p(t) = \delta(t)$, see (13). The correlator efficiency depends strongly on the shape of the signal s(t) and its relationship to h(t) and p(t). The efficiency e_c can typically range from -6 to -2 dB for simple rectangular templates, see (12), to a RAKE gain of several decibels for a delayed reference receiver.

4.4. A UWB Link Budget

The UWB link specifies a transmitter and antenna providing an EIRP of $P_{\rm Tx}$, a receiver with sensitivity $S_{\rm Rx}$, and because the waveform changes shape in the link, a propagation factor P_L determined at a convenient distance. A transmitter operating with about 2 dB margin to a -41.25-dBm/MHz limit over an equivalent bandwidth of 1.3 GHz emits $P_{\rm Tx} = -12$ dBm. A companion UWB receiver operating in AWGN (-174 dBm/Hz), referenced to a data bandwidth W Hz, with noise figure, implementation loss and margin totaling L dB, and operating at an SNR dB signal-to-noise ratio per bit, has a sensitivity of

$$S_{\rm Rx} = -174 + 10\log(W) + L + \rm SNR \, dBm$$
 (14)

The propagation term P_L is evaluated conveniently at one meter using Eq. (8) and the resulting system gain SG at a one meter distance including a receiver antenna gain of G_{Rx} dBi is

$$SG = P_{Tx} - S_{Rx} + P_L + G_{Rx} dB$$
(15)

When $f_c = 4$ GHz, then $P_L = -44$ dB, and using a data bandwidth of W = 40 MHz, losses L = 10 dB, SNR = 7 dB, and $G_{\text{Rx}} = 5$ dBi, the receiver sensitivity is $S_{\text{Rx}} = -69$ dBm, and the system gain at one meter is SG = 30 dB.

Referring to Fig. 11, a 30-dB system gain permits a median range of approximately 10.8 m without RAKE gain. From Fig. 12, the "perfect RAKE" receiver would permit a median range of more than 31 m. Practical implementations would result in a range performance between 10 and 30 m at a 40-Mbps throughput data rate.

5. APPLICATIONS OF UWB

UWB technology uniquely harnesses an ultrawideband of spectrum to provide high bandwidth communications, but in certain implementations also enables indoor precision tracking and radar sensing on top of communications. The unique capabilities of UWB driving it into applications spaces markets include

- High Spatial Capacity. The number of impulses that can be discerned in time over the propagation distance ultimately limits UWB channelization and bandwidth per square meter. For example, with 0.25ns impulses, the upper limit on pulse rate is 4 billion pulses per second. Because of low emitted power ranges are confined to several tens of meters, thus providing exceptional spatial capacities.
- High Channel Capacity and Scalability. Scalability accommodates various channel profiles to harness
the desired data rate given a channel impulse response. Phase coding of impulses simultaneously integrates impulses to improve the energy per data bit, and can RAKE energy from multipath delayed signal echoes.

- Robust Multipath Performance. Multipath signal echoes can be RAKE-received for superior performance indoors. In the limit, the total impulse power on average propagates with an inverse square law just as in free space, in contrast to conventional narrowband harmonic wave radios which tend to propagate more nearly like inverse 3rd power indoors.
- Very Low Transmit Power. Submilliwatt power levels spread over several gigahertz of bandwidth means that the UWB signals will not cause harmful interference to current users of the spectrum, and also will generally be stealthy and less susceptible to detection.
- High System Link Rate. Data rates can be from a high in the hundreds of Mbps, a goal of communications standards developers [19], down to hundreds of kbps. Given a fixed power level, the data rate may be traded off for additional range.
- Location Awareness and Tracking. Some implementations of UWB signaling inherently provide 3D sensing and tracking at centimeter accuracies [17].

5.1. The Role of UWB in Wireless Markets

Advancements in UWB radio technology promise the opportunity of creating unique solutions meeting emerging market needs. There are essentially three basic market spaces in which UWB plays a role:

1. Wireless Communications. As available bandwidth to users increases, applications will continue to evolve to fill the available bandwidth and demand further increases. On top of this increasing demand for bandwidth, the increase in mobile telephony and travel has spurred demand for bandwidth mobility, implying wireless technology. Initial applications of UWB will evolve from the existing market needs for higher speed data transmission, but demand for multimedia-capable wireless is already driving multiple initiatives in the wireless standards bodies. UWB solutions will emerge that are tailored for these applications because of the available high bandwidth. In particular, high-density multimedia applications, such as multimedia streaming in "hotspots" such as airports or shopping centers or even in multidwelling units, will require bandwidths not currently enabled by continuous-wave "narrowband" technologies. The ability to tightly pack high bandwidth UWB "cells" into these areas without degrading performance will further drive the development of UWB solutions. Full-duplex and simplex radio systems using submilliwatt power levels have already been demonstrated with data rates from 78 kbps to hundreds of Mbps at useful ranges in home and office environments.

2. *Precision Tracking*. As the mobility of people and objects increases, up-to-date and precise information about their location becomes a relevant market need. While GPS

and some E911 technologies promise to deliver some level of accuracy outdoors, current indoor tracking technologies remain relatively scarce and have accuracies on the order of 3–10 m. UWB implementations are an adjunct to GPS and E911 that allow the precise determination of location and the tracking of moving objects within an indoor space to an accuracy of a few centimeters. This in turn enables the delivery of location-specific content and information to individuals on the move, and the tracking of high-value assets for security and efficient utilization. While this is an emerging market segment, the accuracy provided by UWB will accelerate market growth and the development of new applications in this area.

3. Radar. Finally, UWB signals enable inexpensive short range high-definition radar. With the new radar capability created by the addition of UWB, the radar market will grow dramatically and radar will be used in areas currently unthinkable. Some of the key new radar applications where UWB is likely to have a strong impact include automotive sensors, collision avoidance sensors, smart airbags, intelligent highway initiatives, personal security sensors, precision surveying, and through-thewall public safety applications. Through-wall radar is already being tested to assist law enforcement and public safety personnel in clearing and securing buildings more quickly and with less risk by providing the capability to detect human presence and movement through walls. Radar enhanced security domes based on precision radar have already demonstrated the capability to detect motion near protected areas, such as high value assets, personnel, or restricted areas. The dome is software configurable to detect movement passing through the edge of the dome, but can disregard movement within or beyond the dome edge.

5.2. Addressing the Wireless Spectrum Squeeze with UWB

UWB operates at ultra-low-power, transmitting impulses over multiple gigahertz of bandwidth. Each pulse, or pulse sequence, is pseudorandomly modulated, thus appearing as "white noise" in the "noise floor" of other radiofrequency devices. UWB operates with emission levels commensurate with levels of unintentional emissions from common digital devices such laptop computers and pocket calculators. Today we have a "spectrum drought" in which there is a finite amount of available spectrum, yet there is a rapidly increasing demand for spectrum to accommodate new commercial wireless services. Even the defense community continues to find itself defending its spectrum allocations from the competing demands of commercial users and other government users. UWB exhibits incredible spectral efficiency that takes advantage of underutilized spectrum, effectively creating a "new" spectrum for existing and future services by making productive use of what appears as the "noise floor" in conventional receiver bandwidths. UWB technology represents a win-win innovation that makes available a critical spectrum to government, public safety, and commercial users.

The best applications for UWB are for indoor use in high-clutter environments. UWB products for the commercial market will make use of the most recent technological advancements in receiver design and will transmit at very low power (submilliwatts). UWB technology enables not only communications devices but also positioning capabilities of exceptional performance. The fusion of positioning and data capabilities in a single technology opens the door to exciting and new technological developments.

BIOGRAPHIES

Kazimierz "Kai" Siwiak received his B.S.E.E. and M.S.E.E. degrees from the Polytechnic Institute of Brooklyn and his Ph.D. from Florida Atlantic University, Boca Raton, Florida. He designed radomes and phased array antennas at Raytheon before joining Motorola, where he received the Dan Noble Fellow Award for his research in antennas, propagation, and advanced communications systems. In 2000, he joined Time Domain Corporation to lead strategic technology development. He has lectured and published internationally; and holds more than 70 patents worldwide, including 31 issued in the United States. He was awarded Paper of the Year by IEEE-VTS and has authored, Radiowave Propagation and Antennas for Personal Communications, (Artech House), now in second edition, and contributed chapters to several other books and encyclopedias.

Laura L. Huckabee received her B.A. degree in physical chemistry in 1986 from Princeton University, her M.A. in comparative culture in 1993 from Sophia University, Tokyo, Japan, and her M.B.A from INSEAD, Fontainebleau, France, in 1994. Her technical work focused on laser spectroscopy and solid state physics. She consulted with U.S. and European firms trying to enter the Japanese market from 1987 through 1993 and joined Procter and Gamble in 1995. At P&G, she developed new businesses in Asia and the Middle East, and moved to the Coca-Cola Company in 1997, developing the Polish soft drink business. In 2000, she returned to the United States to join Time Domain Corporation, the pioneer in UWB radio chipsets, leading international business development and strategic partners.

BIBLIOGRAPHY

- K. Siwiak and L. L. Huckabee, An introduction to ultra-wide band wireless technology, in B. Bing, ed., Wireless Local Area Networks—the New Wireless Revolution, Wiley, New York, 2001.
- R. A. Scholtz and M. Z. Win, Impulse radio, Proc. IEEE Personal, Indoor and Mobile Radio Communications, PIMRC 1997, Helsinki, Finland, 1997.
- M. Z. Win and R. A. Scholtz, Impulse radio: How it works, IEEE Commun. Lett. 2(1): (Jan. 1998).
- 4. Time Domain Corporation, Huntsville, AL (online): <http:// www.timedomain.com> (Nov. 5, 2001).
- 5. K. Siwiak, Ultra-wide band radio: Introducing a new technology, invited plenary paper, *Proc. IEEE Vehicular Technology Conf. 2001*, Rhodes, Greece, May 2001.
- H. L. Bertoni, L. Carin, and L. B. Felson, eds., Ultra-Wideband Short-Pulse Electromagnetics, Plenum Press, New York, 1993.

- L. Carin and L. B. Felson, eds., Ultra-Wideband Short-Pulse Electromagnetics, Vol. 2, Plenum Press, New York, 1995.
- C. E. Baum, L. Carin, and A. P. Stone, eds., *Ultra-Wideband Short-Pulse Electromagnetics*, Vol. 3, Plenum Press, New York, 1997.
- E. Heyman, B. Mandelbaum, and J. Shiloh, eds., Ultra-Wideband Short-Pulse Electromagnetics, Vol. 4, Plenum Press, New York, 1999.
- M. Welborn, System considerations for ultra-wideband wireless networks, *Proc. IEEE Radio and Wireless Conf., RAW-CON 2001*, Boston, MA, Aug. 19–22, 2001, pp. 5–8.
- H. G. Schantz and L. Fullerton, The diamond dipole: A Gaussian impulse antenna, *IEEE APS Conf.*, Boston, July 2001.
- K. Siwiak, T. M. Babij, and Z. Yang, FDTD simulations of ultra-wideband impulse transmissions, *Proc. IEEE Radio* and Wireless Conf., RAWCON 2001, Boston, Aug. 19–22, 2001.
- J. G. Proakis, Digital Communications, McGraw-Hill, New York, 1983.
- R. Hoctor, Transmitted-reference, delay-hopped Ultra-Wideband Communications, Forum on Ultra-Wide Band, Hillsboro, OR (online): http://www.ieee.or.com/IEEEProgram Committee /uwb/uwb.html> (Oct. 11–12, 2001).
- K. Siwiak and A. Petroff, A path link model for ultra-wide band pulse transmissions, *Proc. IEEE Vehicular Technology Conf. 2001*, Rhodes, Greece, May 2001.
- J. Foerster, The effects of multipath interference on UWB performance in an indoor wireless channel, *Proc. IEEE Vehicular Technology Conf. 2001*, Rhodes, Greece, May 2001.
- 17. K. Siwiak, The future of UWB—a fusion of high capacity wireless with precision tracking, Forum on Ultra-Wide Band, Hillsboro, OR (online): <<u>http://www.ieee.or.com/IEEEProgramCommittee/uwb/uwb.html</u>> (Oct. 11–12, 2001).
- K. Siwiak, Radiowave Propagation and Antennas for Personal Communications, 2nd ed., Artech House, Norwood, MA, 1998.
- IEEE P802.15.3, High Rate (HR) Task Group (TG3) for Wireless Personal Area Networks (WPANs), (online): <http://grouper.ieee.org/groups/802/15/pub/TG3.html> (Dec. 1, 2001).

UNEQUAL ERROR PROTECTION CODES

ARDA AKSU North Carolina State University Raleigh, North Carolina

1. INTRODUCTION AND THEORY

The error correcting capability of the majority of algebraic codes is described in terms of correcting errors in codewords, rather than correcting errors in individual digits. However, in many applications some message positions are more important than others. For instance, in transmitting numerical data, errors in the high-order digits are more serious than errors in the low-order digits. Therefore each block of data can be partitioned into classes of different importance (i.e., of different sensitivity to errors). It is possible for a linear block code to provide more protection for selected positions in the input message words than is guaranteed by the minimum distance of the code. Then, it is apparent that the best coding strategy aims at achieving lower bit error rate (BER) levels for more important information bits while admitting higher BER levels for the less important ones. This feature is referred to as unequal error protection (UEP) and linear codes having this property are called *linear unequal error protecting* (LUEP) codes.

UEP codes were first studied by Masnick and Wolf in 1967 [1]. Since then, there has been a proliferation of studies in the field of unequal error protection codes, in both the theory and practical applications. Later work [2–7] investigated various approaches to the construction of UEP codes. A separation vector was later [3], introduced as a measure of the error correcting capability of an UEP code at various locations. For a binary linear (n, k) code C, with generator matrix G, the separation vector $\underline{s} = \{s_1, s_2, \dots, s_k\}$ is defined by

$$s_i = \min\{w(\mathbf{m}G) \mid \mathbf{m} \in \{0, 1\}^k, m_i \neq 0\} \quad i = 1, \dots, k$$

where $w(\cdot)$ denotes the Hamming weight of the argument, namely, the number of nonzero components in the argument.

Another way of looking at it is that the sets $\{\underline{\mathbf{m}}G \mid \underline{\mathbf{m}} \in \{0, 1\}^k, m_i = 0\}$ and $\{\underline{\mathbf{m}}G \mid \underline{\mathbf{m}} \in \{0, 1\}^k, m_i = 1\}$ are at a Hamming distance s_i apart. Hence, for a linear binary (n, k) code C that uses a matrix G for its encoding, complete nearest-neighbor decoding guarantees the correct interpretation of the *i*th digit whenever the error pattern has a Hamming weight less than or equal to $\lfloor (s_i - 1)/2 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer contained in x. From this, it is immediately clear that the minimum distance of the code is $d_{\min} = \min\{s_i \mid i = 1, \ldots, k\}$. Therefore, if a linear code C has a generator matrix G such that the components of the separation vector are not equal, then the code C is called a *linear unequal error protecting* (LUEP) code.

Usually, decoding algorithms for UEP codes are complicated. It has been shown [1,4] that a modified syndrome decoding method using a standard array can be implemented for UEP codes. Essentially, if efficient decoding procedures are available for component codes, then the resulting UEP code can be decoded too. But it is still necessary to design UEP codes, which can be implemented easily.

A binary cyclic code C(n, k) is the direct sum of a number of ideals in the residue class ring $GF(2)[x]/(x^n - 1)$ (where GF is the Galois field) of polynomials in x. In [4], it is shown that an ordering M_1, M_2, \ldots, M_v of generator matrices of these ideals exists such that

$$G = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_v \end{bmatrix}$$

is an optimal generator matrix. The *i*th and *j*th components of the separation vector (<u>s</u>) are equal if the *i*th and *j*th rows of *G* are in the same ideal of $GF(2)[x]/(x^n - 1)$.

If the weight of the generator polynomial of a cyclic code *C* equals $d_{\rm min}$ of the code, then all components of the separation vector are equal. If this is not the case, the separation vector of a cyclic code can be computed by comparing the weight distributions of its cyclic subcodes. In Van Gils [4], compiled a table for UEP capabilities of binary cyclic codes of odd lengths up to 39. Later, Lin et al. [7], extended the table to lengths up to 65 by using exhaustive computer search.

Many of the best known codes can be constructed as generalized concatenated (GC) codes. The construction of these codes use outer codes with different lengths and inner codes (in the columns of the code matrix) with different lengths and distances. The inner code is multiply partitioned, and this partitioning into subcodes is protected by different outer codes. With this method, a large class of optimal linear UEP codes can be generated ([17]). These codes also contain most of the constructions found in van Gils' 1984 paper [6].

Others have investigated coding and decoding schemes to achieve UEP using several convolutional codes with different error correcting capabilities [18-20]. Lower bounds on the free distance of convolutional codes with unequal information protection have been investigated [21,22]. The asymptotic behaviors of these bounds indicate that more gains can be attained for the important data by enlarging the corresponding constraint length. This comes at the cost of reduced performance for the less significant data.

It is desirable to design UEP codes, which can be implemented easily. UEP, in conjunction with modulation, can be achieved by employing either time-division coded modulation (TDCM) or superposition coded modula*tion* (SCM) [8-12]. TDCM is a form of resource sharing in which bit streams of differing importance are transmitted in disjoint modulation intervals. In SCM, the different bit streams are transmitted in the same modulation intervals. Let B_1 and B_2 denote, in decreasing order of importance, the two bit streams to be unequally protected against channel noise, and let r_1 and r_2 denote their respective rates. The rates are given in terms of the bit rate normalized by the total number of modulation intervals available for transmission of these bit streams. To specify unequal error protection requirements, let N_1 and N_2 denote the variances of the Gaussian noise that the respective bit streams need to withstand, where $N_1 > N_2$. In other words, as long as the variance of the channel noise is less than N_i , the bit stream B_i can be decoded with a bit error rate below some prescribed value.

TDCM is a scheme in which the bit streams B_1 and B_2 are transmitted on distinct modulation intervals. Bit stream B_i is transmitted over the fraction α_i of the available modulation intervals (where $\alpha_1 + \alpha_2 = 1$) using channel code C_i and transmission energy e_i . The design problem is usually to select the parameters (α_1, α_2) and (e_1, e_2) such that desired levels of UEP are achieved while minimizing the average transmission energy per modulation interval $e_T = \alpha_1 e_1 + \alpha_2 e_2$. Figure 1 shows a generalized transceiver structure for TDCM.

 $Superposition \ \ coded \ \ modulation \ (SCM) \ \ consists \ \ of transmitting \ both \ bit \ streams \ on \ \ all \ the \ \ available$



Figure 1. Generalized transceiver structure for TDCM.

modulation intervals using a superposition of channel codes in the modulation space. Let us choose constellations S_1 and S_2 with average energies e_1 and e_2 , respectively. Bit stream B_i is encoded with channel code C_i and transmitted using constellation S_i . More specifically, the code C_i generates the codeword \underline{x}_i , which is a sequence of signal points $\{x_i\}$, where $x_i \in S_i$. The codewords \underline{x}_1 and \underline{x}_2 are superimposed and transmitted on the channel as $\underline{x} = \underline{x}_1 + \underline{x}_2$. C_1 and S_1 are respectively referred as outer code and outer constellation and C_2 and S_2 , as inner code and inner constellation. At the decoding end, the received sequence is $y = \underline{x} + \underline{n}$, where N is the variance of the Gaussian noise. Then, if $N_2 < N < N_1$, only bit stream B_1 is reliably decoded. If $N < N_2$, then both bit streams are reliably decoded. The SCM design method aims at achieving the prescribed levels of protection for the two bit streams while minimizing the average transmission energy.

For higher rate transmission with multilevel modulation over bandwidth-limited channels, UEP coding can be achieved in the context of combined coding and modulation because of its efficiency compared to timesharing techniques [13,14]. It has been shown [15] that, asymptotically, the superposition coded modulation technique always outperforms the later one. On the basis of this result, much of the earlier work concentrated on the first technique. Other authors [16] show that this theoretical result doesn't always hold for practical channel codes, which do not achieve capacity. In fact, the opposite happens when a ratio, which measures the degree of inequality in protection, is below a critical threshold.

UEP codes have generated much interest since they are increasingly important in various applications, such as visual communication systems, speech communication, storage and computer systems, and satellite communications. The data in such systems are not equally important, especially after source encoding. Coupled with the proliferation of high-speed wireless networks, which present a very challenging channel for data transfer around the globe, the need for highly efficient UEP coding becomes even more important. The rest of this article focuses on examples of the use of UEP codes in various applications.

2. UEP CODES FOR SPEECH CODING

The trend in current and future cellular mobile radio systems is toward using more sophisticated, digital speech coding techniques. However, mobile radio channels are subject to signal fading and interference, which causes significant transmission errors. The design of speech and channel coding is therefore quite challenging, and UEP codes are considered in the literature for improved performance.

The effects of digital transmission errors on a family of variable-rate embedded subband speech coders have been analyzed [23]. Subband coding of speech is a relatively mature form of waveform coding of speech, in which the signal is first divided into a number of subbands, which are then individually encoded. The underlying principle for the coder is that the bit allocation can be weighed so that those subbands with the most important information get most of the bits. The initial subband coders used fixed bit allocations based on the average spectrum of speech. Later on, the idea of dynamically changing the bit allocation based on the energy of each subband was introduced. An example of the block diagram of the transmitter portion of this coder is shown in Fig. 2. The authors show that there is a difference in error sensitivity of four orders of magnitude between the most and the least sensitive bits of the speech coder. As a result, a family of rate-compatible punctured convolutional (RCPC) codes with flexible UEP capabilities, matched to the speech coder, provides significant gains over a wide range of channel signal-to-noise ratios [24].

Perceptual audio coders (PACs) and similar audio compression techniques are inherently packet-oriented. Audio information for a fixed interval (frame) of time is represented by a variable-bit-length packet. Each packet consists of certain control information followed by quantized spectral/subband description of the audio frame. The key idea behind an UEP scheme is that different components of a packet exhibit varying degrees of sensitivity to channel errors. Experimental results on multilevel UEP schemes, which exploit the unequal impact of transmission errors on various audio components, exhibit significant gains and graceful degradation compared to equal error protection [25].

Use of UEP codes have been suggested for speech coding schemes employed in developing (at the time of writing) third-generation (3G) wireless systems [26] and digital audiobroadcasting, mainly because of its superior performance and graceful performance degradation.

Embedded joint source-channel coding schemes of speech also take advantage of UEP by puncturing symbols of the "RCPC Trellis" encoders [27]. The coder is claimed to be robust to acoustic noise and produce good quality speech for a wide range of channel conditions.

3. UEP CODES FOR IMAGE TRANSMISSION

Robust image transmission has received increasing attention with the advances in wireless communication and image coding. Numerous schemes have been developed for transmission of images over noisy channels.



Figure 2. An example of a dynamic bit allocation subband coder.

One such method employs selective channel coding and transmission energy allocation in conjunction with sequence maximum a posteriori soft-decision detection [28]. An UEP scheme is proposed for transmitting discrete cosine transform (DCT) compressed images over an additive white Gaussian noise (AWGN) channel. The system consists of a TCM scheme that uses selective transmission energy allocation to the DCT coefficients. It also employs a sequence maximum a posteriori (MAP) detection scheme that exploits both channel soft-decision information and the statistical image characteristics. Experimental results indicate that this scheme provides substantial objective and subjective improvements over equal-error protection systems as well as graceful performance degradation. Fei and Ko [29] have used UEP codes for JPEG compressed images in a wireless environment in order to provide extra protection for the header syntax of JPEG compressed images.

Rate-distortion-based error control schemes are also used widely in wireless image communications. Although error correction codes can effectively protect transmitted bits, the introduced overhead bits decrease the available channel capacity. Among various compression techniques for wireless channel transmission, embedded wavelet coders possess excellent rate distortion performance as well as progressive representation property in the transmitted bit stream. In other words, the quality of the decoded images at the receiver end progressively improves as more bits are received. Furthermore, the same number of bits received at the receiver end inherit different capabilities to increase the quality of decoded images due to the different importance of bits at different locations. Therefore, UEP coding schemes based on a rate-distortion model of embedded wavelet image coders generally provide better performance than do equal-errorprotecting schemes [30,31].

Many other schemes investigate the design of the bit allocation between the source coder and channel coder by jointly considering effects of quantization errors and channel errors. Individual bits show different importance for the image reconstruction and different sensitivity to channel errors. Many joint source-channel coding techniques use "subbandwise" UEP, in which the parity bit budget is allocated among subbands proportional to the contained energy [32–34]. This scheme is not optimal since coefficients with large magnitudes at finer scales have more contribution to the distortion reduction than coefficients with small magnitude at coarser scales. Thus "bitplanewise" transmission with UEP is also frequently used in advanced wavelet image codecs [35].

4. UEP CODES FOR VIDEO TRANSMISSION

The standard video compression algorithms (e.g., H.263, MPEG) use predictive coding of frames and variablelength codewords to obtain a large amount of compression. This renders the compressed video bit stream sensitive to channel errors, as predictive coding causes errors in the reconstructed video to propagate in time to future frames of video, and the variable-length codewords cause the decoder to easily lose synchronization with the encoder in the presence of bit errors. To make the compressed bit stream more robust to channel errors, the MPEG-4 video compression standard incorporated several error resilience tools (resynchronization markers, header extension codes, data partitioning, etc.) to enable detection and containment of errors.

However, since wireless channels present quite harsh conditions, powerful error control coding is always needed, and UEP codes present a very good match. The output of an MPEG-4 typical video encoder is a bit stream that contains video packets. These video packets begin with a header, which is followed by the motion information, the texture information, and the stuffing bits (see Fig. 3). The header of each packet begins with a resynchronization

RS	Header	Motion (MVs)	Texture (DCT coefficients)	SB

RS: Resynchronization marker SB: Stuffing bits

Figure 3. An example of a video packet from an MPEG-4 encoder.

marker, which is followed by the important information needed to decode the data bits in the packet. This is the most important information, since the whole packet will be dropped if the header is received in error. The motion information has the next level of information, as motion compensation cannot be performed without it. The texture information is the least important of the four segments of the video packet. Without the texture information, motion compensation concealment can be performed without too much degradation of the reconstructed picture. The stuffing information at the end of the packet has the same priority as the header bits because reversible decoding cannot be performed if this information is corrupted and the following packet may be dropped if the stuffing bits are received in error. When using UEP, the header and stuffing bits normally get the highest amount of protection, the motion bits get the next highest level of protection, and the texture bits receive the lowest level of protection. Using this system, the errors are less likely to occur in the important sections of the video packet. A number of studies are done in order to investigate the performance gains of UEP codes applied to video coding [36,37]. A comparison between using a fixed rate $\frac{7}{10}$ convolutional code for the whole packet or UEP using rate- $\frac{3}{5}$ convolutional code for the header and stuffing segments, a rate- $\frac{2}{3}$ convolutional code for the motion segment, and a rate- $\frac{3}{4}$ convolutional code for the texture segment, shows gains as much as 1 dB in video quality.

Burlina and Alajaji discuss the UEP capabilities of convolutional codes belonging to the family of RCPC codes applied to image coding [38]. The H.263-compatible datastream is partitioned into classes of different sensitivity, without any modification to the standard. Experimental results show that UEP coding provides very good performance when video is transmitted over channels having high round-trip delay and limited bandwidth [39,40].

UEP coding is a very attractive technique in broadcast systems because it gradually reduces the transmission rate. The use of UEP codes for satellite broadcasting of digital TV signals has been considered in [41,42]. The coding scheme is designed in such a way that the information bits carrying the basic definition TV signal have a lower error rate than the high-definition information bits. Because of the nonlinear nature of the channel, the constellations are also chosen to have constant envelope. Reported results achieve graceful degradation, and no error propagation from the first level decoder to the second-level decoder.

5. UEP CODES FOR WIRELESS COMMUNICATIONS

Rapidly developing standards, as well as the technological advancement in wireless communications has enabled the wireless transmission of not only low-quality voice but also high-fidelity multimedia data.

One of the most important performance measures in wireless communication system design is the bit error rate. While this is a meaningful measure for computer applications such as data transfer, it doesn't necessarily reflect the perceptual quality of multimedia data. In particular, for highly compressed audio or video signals, a rare bit error can lead to a highly annoying or unacceptable perceptual distortion. Instead of demanding a very low bit error rate, which is costly to achieve, the use of UEP and joint design of source and channel coders is desirable. Because of this, employing UEP codes in wireless systems have been studied in detail [43,44].

Jung et al. [45], propose a technique for the H.263compatible video datastream, based on the data partitioning technique. The proposed algorithm employs bit rearrangement, which provides the UEP against channel errors. The unequal error protection capabilities of convolutional codes belonging to the family of RCPC codes are presented in many studies [46–48]. RCPC codes can be decoded by the maximum likelihood Viterbi algorithm with full channel state information and soft decisions; hence they are very desirable for use in digital mobile radio channels. Moreover, they are adopted in International Telecommunication Union (ITU) standards H.223 and H.324.

Orthogonal frequency-division multiplexing (OFDM), a type of multicarrier modulation, is the standard modulation scheme for digital audiobroadcasting, and is also proposed for digital television and beyond third generation (3G) cellular communication systems. OFDM can be used naturally and effectively to provide UEP by properly allocating power and assigning a constellation to each individual subchannel. The principle and algorithm of using power allocation is studied in Ho's paper [49].

It has been discovered that the use of soft bits in source decoding of audio, image, and video is beneficial. "Soft" means that the channel or the channel decoder supplies not only binary decisions but also reliabilities to the source decoder. This requires the "soft-in/soft-out" decoders that became available in the late 1990s, as they are used in the powerful "Turbo" decoding schemes. Furthermore, Turbodetection is a suitable method to improve the bit error rate of not only protected bits but also of bits transmitted uncoded in a system with UEP, such as the class 2 bits in the GSM speech channel [50,51].

Figure 4 shows the GSM speech channel at full rate. The RPE-LTP (regular pulse excited-long-term prediction) speech encoder generates a block of 260 bits per 20-ms speech. According to their function and importance for the quality of the speech, the bits of one block are divided into three classes. The most important bits are the class 1a bits (50 bits), which describe the filter coefficients, block amplitudes, and the LTP parameters. Next in importance are the class 1b bits (132 bits), which consist of RPE pointers, RPE pulses, and some other LTP parameters. Least important are the class 2 bits (78 bits), which contain the RPE pulse and filter parameters. First the class 1a bits are encoded by a weak error detecting (53,50,2) CRC (cyclic redundancy check) code. The class 1a and 1b bits are then protected by a convolutional code whereas class 2 bits are transmitted unprotected. The combined sequences of data classes are modulated with the same Euclidean distance properties. As an alternative to the multiplexing of two independent codes, one may employ explicit UEP codes. Then the task of providing unequal protection is divided between the channel encoder and a nonuniform signal set, which discriminates in



Figure 4. Encoding scheme for the GSM speech channel (full-rate TCH/FS).

favor of the more important bits. Sajadieh et al. [52] show that with the above modifications, better bit-error-rate performance for class 1a bits as well as reduced complexity for the overall decoder can be attained.

The wireless interfaces of asynchronous transfer mode (ATM) networks are gaining importance in order to provide mobility. In a wireless ATM, the ATM cells are transmitted via radioframes between a base station and a mobile station. The wireless interface of ATM networks is a band-limited channel with much higher error rate than the wired one. In multimedia wireless ATM interface, the header and various payloads to be transmitted have different importance or error protection needs. Therefore, it is desirable that channel coding provides UEP, matched to the different sensitivity of source encoded symbols. Several papers have been published for studying UEP block codes [53] and convolutional codes [54] in wireless ATMs.

6. MISCELLANEOUS USE OF UEP AND CONCLUDING REMARKS

Asymmetric digital subscriber lines (ADSLs), a transmission system capable of realizing very high bit-rate services over existing telephone lines, are well suited to multimedia applications, such as videotelephony and videoconferencing. Zheng and Liu have shown [55] that UEP codes can be used to provide extra protection for some subchannels, and proposed a method that achieves significant performance improvement compared to spectrally shaped channels commonly used in ADSL systems.

Another interesting study proposes using a coded modulation technique to increase the storage capacity of multilevel and analog memory cells by one extra bit per cell [56]. The technique is applied to improve readout bit -error probability and provide unequal error protection for the bits to be stored.

As can be seen from the previous examples, UEP codes have been applied to many aspects of modern communication systems when there is a natural need to protect some part of the information content more than the others. It is also not too difficult to see that many data sources fall into this category. Throughout the published literature on UEP codes, some commonalities can be observed: (1) if designed properly, UEP codes generally provide better performance than can equal error

protection codes of the same category, and (2), UEP codes achieve gradual and smooth degradation of performance. And because of these properties, UEP codes are employed in various types of communication networks.

BIOGRAPHY

Arda Aksu received his B.Sc. degree from Middle East Technical University, Ankara, Turkey, in 1989, and his M.Sc. and Ph.D. degrees all in electrical engineering from Northeastern University, Boston, Massachusetts, in 1992 and 1995, respectively. From 1994 to 2000 he was with Verizon Technology Organization (formerly known as GTE Laboratories Inc.) in Waltham, Massachusetts. His responsibilities as a principal member of technical staff mainly included design, analysis, and testing of cellular and fixed-access wireless communications systems. Since 2001, he has been a senior technologist with the office of the CTO in Comverse working on the conceptualization and development of new and enhanced services for the next generation wireless networks.

Dr. Aksu's research interests are in the area of technology assessment and strategy development for the provision of wireless and personal communication services and products, as well as the analysis, design, and implementation of wireless networks. He is the author of over 25 scientific and technical publications as well as industry standards contributions in the above disciplines. He has served in the capacity of session chair, paper reviewer, and invited author in IEEE communities. He is the coinventor in five U.S. patent filings.

BIBLIOGRAPHY

- B. Masnick and J. K. Wolf, On linear unequal error protection codes, *IEEE Trans. Inform. Theory* **IT-3**: 600–607 (Oct. 1967).
- 2. W. C. Gore and C. C. Kilgus, Cyclic codes with unequal error protection, *IEEE Trans. Inform. Theory* (March 1971).
- I. M. Boyarinov and G. L. Katsman, Linear unequal error protection codes, *IEEE Trans. Inform. Theory* **IT-27**: 168–175 (March 1981).
- W. J. van Gils, Two topics on linear unequal error protection codes: Bounds on their length and cyclic code classes, *IEEE Trans. Inform. Theory* **IT-29**: 866–876 (Nov. 1983).

2768 UNEQUAL ERROR PROTECTION CODES

- L. A. Dunning and W. E. Robbins, Optimal encoding of linear block codes for unequal error protection, *Inform. Control* 37: 150-177 (1978).
- W. J. van Gils, Linear unequal error protection codes from shorter codes, *IEEE Trans. Inform. Theory* **IT-30**: 544–546 (May 1984).
- M. C. Lin, C. C. Lin, and S. Lin, Computer search for binary cyclic UEP codes of odd length up to 65, *IEEE Trans. Inform. Theory* 36: 924–935 (July 1990).
- C. Zhi, F. Pingzhi and J. Fan, New results on self-orthogonal unequal error protection codes, *IEEE Trans. Inform. Theory* 36: 1141–1144 (Sept. 1990).
- A. R. Calderbank and N. Seshadri, Multilevel codes for unequal error protection, *IEEE Trans. Inform. Theory* 39: 1234–1248 (July 1993).
- L.-F. Wei, Coded modulation with unequal error protection, IEEE Trans. Commun. 41: 1439-1449 (Oct. 1993).
- A. Aksu and M. Salehi, Transmission of analog sources using unequal error protecting codes, *IEEE Trans. Commun.* 43: 1225–1229 (Feb./March/April 1995).
- N. Seshadri and C.-E. W. Sundberg, Multilevel block coded modulations for the Rayleigh fading channel, *Proc. IEEE Global Commun. Conf.*, Phoenix, AZ, 1991.
- M. Isaka and H. Imai, Hierarchical coding based on adaptive multilevel bit-interleaved channels, *Proc. IEEE Vehic. Tech. Conf.*, 2000, pp. 2227–2231.
- M. Isaka et al., Multilevel codes and multistage decoding for unequal error protection, *Proc. IEEE Int. Personal Wireless Commun. Conf.*, 1999, pp. 249-254.
- P. P. Bergmans and T. M. Cover, Cooperative broadcasting, IEEE Trans. Inform. Theory 20: 317–324 (May 1974).
- S. Gadkari and K. Rose, Time-division versus superposition coded modulation schemes for unequal error protection, *IEEE Trans. Commun.* 47: 370-379 (March 1999).
- U. Dettmar, G. Yan and U. K. Sorger, Modified generalized concatenated codes and their application to the construction and decoding of LUEP codes, *IEEE Trans. Inform. Theory* **41**: 1499–1503 (Sept. 1995).
- M. Matsunaga, D. K. Asano, and R. Kohno, Unequal error protection scheme using several convolutional codes, *Proc. IEEE Int. Symp. Inform. Theory*, 1997, p. 101.
- M.-C. Chiu, C.-C. Chao, and C.-H. Wang, Convolutional codes for unequal error protection, *Proc. IEEE Int. Symp. Inform. Theory*, 1997, p. 290.
- C.-H. Wang and C.-C. Chao, Further results on unequal error protection of convolutional codes, *Proc. IEEE Int. Symp. Inform. Theory*, 2000, p. 35.
- T. Danon and S. I. Bross, Free distance lower bounds for unequal error-protection convolutional codes, *Proc. IEEE Int. Symp. Inform. Theory*, 2000, pp. 261.
- E. K. Englund, Nonlinear unequal error-protection codes are sometimes better than linear ones, *IEEE Trans. Inform. Theory* **IT-37**(5): 1418–1420 (Sept. 1991).
- R. V. Cox, J. Hagenauer, N. Seshadri, and C.-E. W. Sundberg, Subband speech coding and matched convolutional channel coding for mobile radio channels, *IEEE Trans. Signal Process.* **39**: 1717–1731 (Aug. 1991).
- 24. H. Shi, P. Ho, and V. Cuperman, Combined speech and channel coding for mobile radio applications, *Proc. IEEE*

Int. Conf. Select Topics Wireless Commun., 1992, pp. 180–183.

- D. Sinha and C.-E. W. Sundberg, Unequal error protection methods for perceptual audio coders, *Proc. IEEE Int. Conf.* Acoustics, Speech, Signal Process., 1999, pp. 2423–2426.
- S. Kim, B. Kang, and K. Oh, An improved scheme in CDMA forward link using unequal error protection, *Proc. IEEE Int. Conf. Universal Personal Commun.*, 1998, pp. 1093–1096.
- A. Bernard, X. Liu, R. Wesel, and A. Alwan, Embedded joint source-channel coding of speech using symbol puncturing of trellis codes, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.* 1999, pp. 2427–2430.
- F. I. Alajaji, S. Al-Semari, and P. Burlina, Visual communication via trellis coding and transmission energy allocation, *IEEE Trans. Commun.* 47: 1722–1728 (Nov. 1999).
- X. Fei and T. Ko, Turbo-codes used for compressed image transmission over Rayleigh fading channel, *Proc. IEEE Int. Conf. Universal Personal Commun.*, 1997, pp. 505-509.
- T.-C. Yang and C.-C. J. Kuo, Error correction for wireless image communication with a rate-distortion model, *Proc. Signals, Systems, Computers Conf. (Asilomar)*, 1998, pp. 963-967.
- I. Kozintsev and K. Ramchandran, Hybrid compressed uncompressed framework for wireless image transmission, *Proc. Int. Conf. Image Processing*, 1997, pp. 77–80.
- P. G. Sherwood and K. Zeger, Error protection for progressive image transmission over memoryless and fading channels, *IEEE Trans. Commun.* 46: 1555–1559 (Dec. 1998).
- M. J. Ruf, A high performance fixed rate compression scheme for still image transmission, *Proc. Data Comput. Conf. (DCC)*, 1994, pp. 294–303.
- S. Manji and G. Djuknic, Bandwidth efficient and error resilient image coding for Rayleigh fading channels, *Proc. IEEE Vehic. Technol. Conf.* (VTC), 1999, pp. 1485–1489.
- J. Vass and X. Zhuang, Joint source-channel coding for highly efficient error resilient image transmission, *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, 2000, pp. 311–314.
- W. R. Heinzelman, M. Budagavi, and R. Talluri, Unequal error protection of MPEG-4 compressed video, *Proc. Int. Conf. Image Processing (ICIP)*, 1999, pp. 530–534.
- T. Kawahara and S. Adachi, Video transmission technology with effective error protection and tough synchronization for wireless channels, *Proc. Int. Conf. Image Processing (ICIP)*, 1996, pp. 101–104.
- P. Burlina and F. Alajaji, An error resilient scheme for image transmission over noisy channel with memory, *IEEE Trans. Image Process.* 593-600 (April 1998).
- 39. A. Andreadis, G. Benelli, A. Garzelli, and S. Susini, FEC coding for H.263 compatible video transmission, *Proc. IEEE Int. Cong. Image Processing*, 1997, pp. 579–581.
- C. W. Yap and K. N. Ngan, Unequal error protection of images over Rayleigh fading channels, *Proc. Int. Symp. Signal Proc. Applications (ISSPA)*, 1999, 19-22.
- R. H. Morelos-Zaragoza, O. Y. Takeshita, and H. Imai, Coded modulation for satellite broadcasting, *Proc. IEEE Global Telecommun. Conf. (GLOBECOMM'96)*, 1996, 31–35.
- 42. G. Taricco and E. Biglieri, Some simple coded-modulation schemes for unequal error protection in satellite communications, *Proc. IEEE Int. Symp. Spread Spectrum Techniques*, *Applications*, 1996, 1288-1294.

- H. Gharavi and S. M. Alamouti, Multi-priority video transmission for third generation wireless communication systems, *Proc. IEEE* 87: 1751–1763 (Oct. 1999).
- 44. F. Babich, G. Lombardi, and F. Vatta, Performance evaluation of source-matched channel coding for mobile communications, *Proc. IEEE Vehic. Technol. Conf.*, 1998, 2517– 2521.
- 45. H.-S. Jung, R.-C. Kim, and S.-U. Lee, On the robust transmission technique for H.263 video data stream over wireless networks, *Proc. IEEE Int. Image Proc. Conf.*, 1998, 463– 466.
- 46. J. Hagenauer, N. Seshadri, and C.-E. Sundberg, The performance of rate-compatible punctured convolutional codes for digital mobile radio, *IEEE Trans. Commun.* 38: 966–980 (July 1990).
- 47. S. Kang and K.-Y. Yoo, Region and time based unequal error protection for video transmission over mobile links, *Proc. IEEE Int. Symp. Circuits and Systems*, 1999, 511–514.
- J. Hagenauer and T. Stockhammer, Channel coding and transmission aspects for wireless multimedia, *Proc. IEEE* 87: 1764-1777 (Oct. 1999).
- K.-P. Ho, Unequal error protection based on OFDM and its application in digital audio transmission, *Proc. IEEE Global Telecommun. Conf. (GLOBECOM'98)*, 1998, 1320–1325.

- G. Bauch and V. Franz, Iterative equalization and decoding for the GSM system, *Proc. IEEE Vehic. Technol. Conf.*, 1998, 2262-2266.
- 51. F. Burkert et al., Turbo decoding with unequal error protection applied to GSM speech coding, *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, 1996, 2044–2048.
- M. Sajadieh, F. R. Kschischang, and A. Leon-Garcia, Modulation-assisted unequal error protection over the fading channel, *IEEE Trans. Vehic. Technol.* 47: 900–908 (Aug. 1998).
- 53. S. Aikawa, Y. Motoyama, and M. Umehira, Forward error correction schemes for wireless ATM, *Proc. IEEE Int. Commun. Conf.*, June 1996, pp. 454–458.
- 54. Z. Sun, S. Kimura and Y. Ebihara, Adaptive two-level unequal error protection convolutional code scheme for wireless ATM networks, *Proc. IEEE INFOCOM*, 2000, pp. 1693-1697.
- H. Zheng and K. J. Ray Liu, Robust image and video transmission over spectrally shaped channels using multicarrier modulation, *IEEE Trans. Multimedia* 88-103 (March 1999).
- H.-L. Lou and C.-E. Sundberg, Coded modulation to increase storage capacity of multilevel memories, *Proc. IEEE Global Telecommun. Conf.*, (*GLOBECOM'98*), 1998, 3379–3384.

V.90 MODEM

BRENT TOWNSHEND Townshend Computer Tools Menlo Park, California

1. INTRODUCTION

Since the era of the earliest time-shared computers, data communications methods have followed a path parallel to computer development. Each new generation of faster CPU and denser memory has been accompanied by improvements in communication capabilities. One particularly important mode of communication has been transmission of data over regular telephone lines. However, since the telephone system was not designed with data transmission in mind, creative manipulations of the signal were required to achieve this. One of the earliest examples of this was the teletype machine, which sent data at a speed of 110 bits per second (bps). This was achieved by coding the bits into one of two different tone signals and sending one such symbol 110 times each second. The tones were chosen to be in the audio range that the telephone system was designed to pass, and adequately separated to allow the receiving end to distinguish which tone was present at each time interval. This pattern of modulating a signal, such as alternating between the two tones above, and the demodulation at the receiver to reconstruct the original data signal was the basic model and namesake for all of the subsequent modems (modulator-demodulators) that followed.

Clearly, the key requirement of a modem was to accurately transmit the data accurately and to do so as fast as possible. Researchers found ways to continually increase the speed by applying more and more complex techniques to the modulation so that more data could be squeezed into the telephone channel designed for voice. The result was a rapid succession of improvements, typically with the transmission rate doubling every 3 years. It would be misleading, however, to simply imagine that these developments were successive improvements of one idea. Each new generation of modem pioneered a qualitatively different technique of modulation or demodulation to achieve the next step in the progression. The 300- and 1200-baud¹ modems based on frequency-shift keying gave way to 2400-bps devices, by exploiting phase shift keying and coherent demodulation. With the introduction of adjustable and then adaptive equalizers, more complex signaling schemes were introduced that could carry more per second in the same bandwidth. Amplitude modulation was combined

with phase shift keying to give the quadrature amplitude modulation allowing rates of 9600, 14,400, and 19,200 bps by the late 1980s.

Viterbi decoding and increasingly complex modulation patterns allowed further optimization to bring the data rates to 33,600 bps by 1994. But at this point it was thought that the maximum had been obtained, and there was little optimism that any further improvement could be made regardless of the ingenuity of the modulation designers.

2. SHANNON LIMIT

Throughout and even before the evolution described above, there was a firm theoretical foundation for the envelope of possibilities that had been developed by Claude Shannon several decades earlier. In his seminal paper of 1948 [1], he provided a simple equation to predict that maximum data rate that can be obtained on any channel given the width of the frequency band that is passed by the channel, W, the signal power, S, and the noise power N:

$$C = W \log_2 \frac{S+N}{N}$$

This agrees with intuitive ideas of how data rate should vary. If the bandwidth is doubled (i.e., equivalent to having two identical channels, each with the original bandwidth), then the data rate is doubled. As the background noise level is lowered or the signal power is increased, each symbol is less "fuzzy" and more symbols can be packed into the available space, again resulting in higher data rates.

Applying the Shannon limit to the telephone channel is straightforward. The telephone channel passes audio frequencies in the range of 300–3400 Hertz, giving a bandwidth of 3100 Hz. AT&T and its predecessors [2] have repeatedly surveyed the noise level of the U.S. telephone channels, finding a consensus signal-to-noise ratio of approximately 35 dB. Using these values in the above equation yields a result of 36,000 bps. Thus, it seemed clear at one time that the 33,600-bps rate attained above leaves little further room for improvement.

3. BREAKING THE LIMIT

As we now know, 56-kbps modems are possible over the same channel analyzed above. Does this mean that the Shannon limit is in error or somehow inapplicable? No, the theory is correct, and even the calculation for the telephone channel is accurate. However, what the analysis presented above overlooks is that the definition of the inputs to the equation depends on your point of view. Specifically, the signal-to-noise ratio depends on your definition of what constitutes noise. For example, imagine tuning a radio between two stations such that

 $^{^1}$ The baud rate is the number of symbols that are transmitted per second — if the symbol conveys one bit, then this is equal to the bit rate.

both are heard. If you are interested in listening to one, say, a jazz station, then you may think of the other, a rock station, as a noise source. Conversely, a different listener may swap the classification of which is the signal and which is the noise.

In the case of the telephone channel, the definition of noise is equally important. In the original telephone system design, signals were carried in their analog form from one endpoint to the other through a series of switches and relays. This system picked up electrical noise from a variety of sources including cross-talk from other calls and inductive pickup from motors and other electrical sources. Eventually, the network was replaced by a digital system where the signals are carried between central offices in a digital form that is immune to electrical noise pickup. In the design of the digital system, a choice had to be made as to the sampling rate and precision of these data. To minimize the data requirements and maximize the system capacity in terms of the number of concurrent voice connections, these choices were made to approximate the quality of the older analog system. Thus, during that evolution of the long-distance network from analog to hybrid to all digital, the electrical noise, which gave a signal noise ratio of around 35 dB, was replaced by quantization noise of the same magnitude. As far as telecom designers were concerned, the system did not look very different-the bandwidth and signal-to-noise ratios were similar and could generally be treated identically. But this rough equivalence ignores a critical difference. Specifically, if you are in control of the digital words used by the telephone network, then the quantization is no longer noise—it is a deterministic phenomenon [3,4].

The insight that the quantization is not noise radically changes the parameters of the problem. The pertinent noise in the Shannon calculation is then the noise due to electrical pickup, crosstalk, and other uncontrollable sources. Surveys of the local loops [2] estimate these noise levels at lower than -79 dBm^2 for at least 90% of the cases. Since telephone companies typically limit transmit levels to -12 dBm, the Shannon limit gives a maximum data capacity of

$$C = 3100 \log_{2} 10^{(79-12)/10} = 69 \text{ kbps}$$

This capacity for the local loop is further demonstrated by the operation of products such as ISDN and DSL, which

 2 The unit dBm is a measure of absolute power in decibels such that 0 dBm is one milliwatt.

use the same local loop to achieve data rates of ≥ 1.5 Mbps, although they use a wider bandwidth.

4. OVERVIEW OF V.90

A V.90 communication system is shown in Fig. 1. The digital modem endpoint usually resides at a commercial server such as an Internet service provider (ISP) and may be part of a remote access server that contains many such devices. The digital modem is connected to the telephone network via ISDN, T1, or E1 digital subscriber lines. These types of connections maintain digital signaling from the modem to the digital-to-analog converter at the digital central office (CO) where the end user's line is terminated.

Digital telephone networks carry both voice and data signals as digital symbols or codewords. Each symbol consists of 8 bits and is sent at a rate of 8000 symbols per second, giving a data rate of 64 kbps. When the symbols represent a voice signal, they are interpreted according to either a μ -law or A-law encoding rule³ [5]. This nonlinear encoding allows the dynamic range and noise characteristics of a typical telephone channel to be preserved when the signal is carried digitally. As can be seen in Fig. 2, these encoding rules allow small signals to be encoded with high precision while still permitting large signals to be transmitted.

Unlike previous telephone modems such as V.34, the V.90 modem uses a baseband signaling scheme without modulation. The data to be transmitted are mapped onto a sequence of telephone codewords and then presented to the telephone network as if they represented an analog signal. At the central office the stream of symbols received from the network must be converted to an analog signal to be sent to the consumer's telephone. The central office does not distinguish between these signals and normal voice calls, so it applies the digital data to a codec, which converts the symbols to an analog signal using the μ -law or A-law rules. The codec also filters the output of the digital-to-analog conversion to smooth the signal and reduce high-frequency components before sending the signal along the two-wire local loop to the consumer.

At the customer premises, an analog modem is installed that is connected to a telephone jack. It receives the audio signal that was sent by the codec after additional filtering by the transmission lines and pickup of noise and

³ The μ -law nonlinear mapping is used primarily in North America, while A law is used in other parts of the world.



Figure 1. V.90 communications path.



Figure 2. Nonlinear mapping of μ -law codewords.

crosstalk. The analog modem must undo those effects as well as the effects of the codec's smoothing filter. It must then reconstruct the sample timing of the codec and infer the sequence of codewords that was transmitted over the network. Once the symbol stream has been reconstructed, the mapping used by the digital modem can be reversed to recover the original data.

In addition, data must be sent back from the analog modem to the digital end. This upstream transmission is usually done using a lower bandwidth modulation technique such as V.34 at 28.8 or 33.6 kbps. This can be done without significantly impacting the downstream data through use of echo cancelers at both ends, which separate the upstream and downstream signals even though both are carried on the same 2-wire local loop. The resulting asymmetric data rates, 56 kbps downstream and 33.6 kbps upstream, are well suited to most traffic patterns. The typical use of Internet access requires significantly more downstream data such as HTML, images, and video, but the upstream data more often consists of short requests, acknowledgments, and so on.

5. DIGITAL MODEM OPERATION

gram.

The internal structure of the digital modem is shown in Fig. 3. It starts with an interface to the host computer or source of data. Since the digital modem usually resides at a central server, such as an Internet service provider (ISP), several, if not hundreds, of digital modems may be implemented by a single hardware device. In any case, the data from the host computer eventually reach the PCM encoder, where the data are is converted into a sequence of 8-bit symbols for presentation to the telephone network.

On the surface, the PCM encoder (Fig. 4) appears to have a simple function; to pack the incoming data onto telephone network codewords. Once encoded, the codewords would be transmitted at a rate of 8000 per second, providing a 64-kbps data path. However, several factors make the operation more complex, including

Codec Filtering. At the receiving central office, the smoothing filter in place for all telephone calls filters low-frequency components out of the reconstructed analog signal. A typical codec (compression/decompression) filter is shown in Fig. 5. As can be seen from the figure, signals below 50 Hz are removed by this filter. Since this filter cannot be bypassed without making changes at the central office, data sequences that result in lowfrequency components would not be uniquely identifiable by the analog modem. To avoid this, the digital







Figure 5. Codec digital-to-analog converter smoothing filter transfer function.

modem sends only a subset of all possible codeword sequences such that the resulting signal has a spectral shape with little energy outside the frequencies passed by the codec.

- Codec Nonlinearities. The digital-to-analog converters used in codecs are far from ideal. Although the analog levels corresponding to each codeword value are specified in the G.711 standard [5], actual implementations may introduce DC biases, nonlinearities, or asymmetries in the mapping. These nonlinearities need to be learned by V.90 modems for each call and appropriate compensation must be applied.
- Local Loop Transmission. The physical pair of wires between a central office and a consumer's home or business is far from an ideal medium. Wiring anomalies, such as bridge taps, create topologies with unterminated branches that cause signal reflections. Loading coils, which are used to improve the frequency response for voice calls, can wreak havoc for data signals. Modems must identify the existence of these and choose appropriate compensation or fallback to lower speeds.
- *Power Constraints.* Arbitrary sequences of codewords may result in signal levels on the local loop that are higher in power than typical telephone signals. This could give rise to crosstalk to adjacent telephone lines causing noise on other calls. To prevent this, the Federal Communications Commission (FCC)



and other regulatory organizations have specified power limits making certain codeword sequences impermissible.

- Digital Impairments. Even though telephone codewords are carried digitally from one central office to another, the network does, at times, modify that datastream. Digital or even analog attenuation is sometimes inserted to control echo levels. This may be done using a remapping of codewords or with a tandem of digital-to-analog and analog-to-digital converters. When central offices need to communicate signaling information, one method commonly used is "robbed bit" signaling [12]. The network will use the least-significant bit of one out of every 6 (or sometimes 12) codewords for its own signaling. During a single call, multiple hops may entail use of multiple codewords for signaling within each 6codeword frame. Since the original value of these bits is lost, fewer bits of each codeword are usable for data transmission. For international calls a remapping between the two G.711 codeword sets (μ -law and A-law) may occur.
- Speech Coding. Some transmissions undergo compression, such as ADPCM, designed to pass speech signals at lower bit rates. This allows more circuits to be carried on the same digital trunks. However, such lossy coding techniques limit the data rates and, in many cases, prevent the use of high speed techniques such as V.90.

For these reasons, the encoder must detect these problems and, if possible, provide some preprocessing of the data stream to avoid them and ensure that the data is recoverable by a decoder.

Figure 4, a block diagram of the PCM Encoder, shows the sequence of these transformations given in the ITU V.90 Recommendation [6]. These have been standardized to guarantee inter-operability between different implementations of the modems.

Generation of codewords is done in frames of 6 codewords to allow repeatable handling of digital impairments such as robbed-bit signaling, which usually occurs in fixed positions within each 6-codeword frame. By creating an encoder that handles each of these 6 symbol intervals separately, it is possible to make use of the least significant bit in intervals where robbed-bit signaling is not occurring and avoid intervals where it does occur.

The first step in packing the incoming data into these 6 codewords is bit selection. As will be discussed below, the sign bit of each symbol is treated specially, so at this point blocks of data bits are separated into S bits that will be encoded into the codewords' sign bits and K bits that will determine the magnitude part of each codeword. For example, at the maximum rate allowed by the standard, groups of 42 bits are read from the interface and separated so that 3 are used for sign bit encoding and the remaining 39 are passed on to the modulus encoder. The data rate using this breakdown is

Rate =
$$\frac{8000}{6} * (K + S)$$

= $\frac{8000}{6} * (39 + 3)$
= 56,000 bps

The standard also allows for other divisions where between 3 and 6 bits are used for sign bit encoding and 15-39 bits are used by the modulus encoder, providing data rates ranging from 28 to 56 kbps.

The next step in the encoding process is using the K bits to choose the magnitude portion of the 6 codewords that will be sent to the telephone network. Each magnitude has 7 bits of precision, which would allow up to 42 bits of encoding. However, only a subset of all possible codewords is used in each symbol interval. These codesets, $\{C_0\}$ to $\{C_5\}$, are determined in the training phase during the initiation of a call (see Section 9 below). Each codeset has M_i elements where $M_i \ll 128$ and the modulus encoding step consists of choosing six integers, K_0 to K_5 , from the six codesets $\{C_0\}$ to $\{C_5\}$. The algorithm for choosing the K_i has the following steps:

1. Use the *K* incoming bits to represent an integer, R_0 :

$$R_0=\sum_{j=0}^{k-1}b_j2^j$$

2. Perform the following recursion to obtain the K_i :

$$K_i = R_i ext{ modulo } M_i$$
 $R_{i+1} = R_i - K_i M_i$

Assuming that the product of the codeset sizes is greater than 2^{K} , these steps result in a set of K_i that can be used to reconstruct the R_0 and the original K data bits:

$$R_0 = \prod_{i=0}^6 k_i$$

For example, if K = 15, S = 4, and, during training, the M_i were chosen as $\{6, 4, 6, 6, 7, 6\}$, then to encode the data bits $\{11010101000100110\}$, we would use the first 4 bits

{1101} for sign encoding. The remaining 15 bits would then be applied to the modulus encoder:

$$\begin{split} R_0 &= \sum_{j=0}^{14} b_j 2^j \\ &= 0.2^0 + 1.2^1 + 0.2^2 + 1.2^3 + 0.2^4 + 0.2^5 \\ &+ 0.2^6 + 1.2^7 + 0.2^8 + 0.2^9 + 1.2^{10} + 0.2^{11} \\ &+ 0.2^{12} + 1.2^{13} + 1.2^{14} = 10,387 \end{split}$$

$$\begin{split} K_0 &= (R_0 \text{ modulo } M_0) \qquad R_1 = \frac{R_i - K_i}{M_i} = 1731 \\ &= 10,387 \text{ modulo } 6 = 1 \end{split}$$

$$\begin{split} K_1 &= 1731 \text{ modulo } 4 = 3 \qquad R_2 = \frac{1731 - 3}{4} = 432 \\ K_2 &= 432 \text{ modulo } 6 = 0 \qquad R_3 = \frac{432 - 0}{6} = 72 \\ K_3 &= 72 \text{ modulo } 6 = 0 \qquad R_4 = \frac{72 - 0}{6} = 12 \\ K_4 &= 12 \text{ modulo } 7 = 5 \qquad R_5 = \frac{12 - 5}{7} = 1 \\ K_5 &= 1 \text{ modulo } 6 = 1 \end{split}$$

The resulting output from the modulus encoder would then be $\{1, 3, 0, 0, 5, 1\}$. These values are then used to index into the codesets, $\{C_0\}$ to $\{C_5\}$ to choose the magnitude portion of each of the six codewords to be transmitted next to the telephone network.

The S sign bits taken from the data are combined with $S_r = 6 - S$ additional bits that are chosen by a spectral shaper. The function of the spectral shaper is to choose the S_r additional bits to control the spectrum of the analog signal that will be constructed by the codec at the consumer's CO. As discussed above, the codec applies its own filtering to the signal. Performance can be improved by avoiding symbol sequences that have significant energy in parts of the spectrum where the codec attenuates the signal.

During training, the characteristics of the codec filter and transmission lines are probed and a spectral shaping filter is chosen by the analog modem of the form

$$F(z) = \frac{(1-b_1z^{-1})(1-b_2z^{-1})}{(1-a_1z^{-1})(1-a_2z^{-1})}$$

The spectral shaper applies this filter to the projected output over the next several frames for each possible choice of the S_r additional sign bits and chooses the values that minimize the energy output from the filter.

In the example above, S = 4 and $S_r = 2$, so the energy output from the spectral filter would be measured using the four possible settings of the two additional sign bits combined with the magnitude bits output from the modulus encoder. The chosen bits are the ones that result in the lowest energy over the current and subsequent frames (the standard allows this window to range from 1 to 4 frames). Note that the V.90 standard includes additional manipulations of the sign bits, but the effect is similar.

6. ANALOG MODEM OPERATION

The analog modem end of a V.90 connection must reverse the effects of the codec at the central office and the distortions caused by the local loop. It must also synchronize itself to the clock used by the telephone company to infer the digital codewords that were sent on the digital telephone network and undo the digital modem transformations to recover the original data signal.

The V.90 standard does not specify how the decoder should operate since its implementation does not impact the interoperability aspects. Given the function of the encoder, any decoder that extracts the original bits is compliant. However, a decoder implementation will typically consist of an equalizer, a clock recovery circuit, decision logic, and bit demapping to reverse the effects of the modulus encoding and spectral shaping. A block diagram of a possible implementation is shown in Fig. 6.

The main elements of the decoder are an equalizer that provides spectral compensation to undo the effects of the codec and local loop filtering; a clock recovery circuit to synchronize the modem with the central office's 8-kHz sampling rate, and decision logic that makes a decision as to which bits were transmitted by the digital modem. In addition, the analog modem will include logic for initialization, training, retraining, compression, error correction, and upstream data transmission as described below.

6.1. Equalizer

The equalizer can be implemented using an adaptive equalizer such as those described by Gitlin et al. [9]. The combination of the codec filter and the local loop has a spectral characteristic, shown in Fig. 5, that attenuates signals below 400 Hz and those above 3400 Hz. It also creates a slight spectral slope and adds some phase distortion to the signal. The equalizer must amplify the low and high frequencies, compensate for the spectral slope, and remove the phase distortion. One possible implementation of an equalizer that achieves these goals is the combination of a fractionally spaced feedforward equalizer with a decision feedback equalizer as shown in Fig. 7. The fractional spacing allows recovery of the high-frequency signals that are near the Nyquist rate of 4 kHz with better performance and lower complexity than sampling at 8 kHz would allow. The decision feedback equalizer permits reconstruction of the low-frequency components where inversion of the nulls of the codec filter using a conventional filter would result in an unstable system. By including the decision logic in the feedback loop, noise is not amplified by the filter as long as the decisions are correct. By choosing the codeword set appropriately, the error rate for the decisions can be made arbitrarily small.

The equalizer will normally be trained and its parameters established during the call initiation. In addition, to track changing characteristics of the channel, adaptation can occur continuously using an internally generated error signal. The tap weights of the equalizer are adjusted to minimize the error estimate using any of a host of update algorithms, such as LMS.

6.2. Clock Recovery

As well as equalization, the analog modem must lock to the codec's 8000-Hz sample clock used by the telephone network without having any direct connection to that clock. During call initiation a known pattern of codewords is transmitted and the analog modem deduces the frequency and phase of the sampling clock from these. In much the same way that the equalizer is adapted to changing channel conditions, the phase-locked clock can also track the codec clock using an error signal.

6.3. Decision Logic

With the equalized signal and a recovered sampling clock, the analog modem can then estimate the analog signal



Figure 6. Analog modem block diagram.



Figure 7. Equalizer structure.

output by the codec at each sampling instant. With the knowledge of the output levels for each of the M_i possible input codewords, the codeword that would give the closed analog signal is chosen to reconstruct K_i . In this way, the codewords that arrived at the codec at each frame are recovered. The operations performed by the digital modem's PCM encoder can then be reversed to reconstruct the K + S data bits for each frame. Assuming the decision of the K_i was correct, this also gives the analog modem information as to the degree of error in the equalizer outputs. By subtracting the theoretical analog value for the received codeword from the equalizer output, an error signal is formed. This signal can be used to control adaptation of the equalizer and clock synchronizer.

7. REVERSE CHANNEL AND ECHO CANCELLATION

The preceding descriptions of the digital and analog modem focus only on the downstream channel from the digital modem to the analog modem. In V.90, the upstream channel is implemented using V.34 modulation techniques, giving a maximum upstream data rate of 33.6 kbps. This asymmetric channel is well suited to most applications where significantly more data flow downstream to the client than upstream.

Introduction of the reverse channel does add significant complexity to the system in that echo cancellation has to be added so that the modems do not receive delayed versions of their own outgoing signals at levels that would corrupt the incoming data.

8. COMPRESSION AND ERROR RECOVERY

Data sent over a modem connection often have some structure that can be exploited by a lossless compressor to improve throughput. HTML, for example, can typically be compressed by a factor of 4 or more. V.90 and prior modem standards often employ a form of Lempel–Ziv–Welch [7] compression to achieve this. Data that are not compressible (such as data that have already been compressed) can be detected using a continuous "compressibility" monitor. The compressor would then be switched off and the data passed verbatim. The details of these compression algorithms can be found in ITU Recommendation V.42bis [8].

Error control is usually implemented using the *Link*access procedure for modems (LAPM), which has been standardized as V.42 [14]. This procedure blocks data together in packets and adds a checksum to each packet. When packets with correct checksums are received, an acknowledgement is returned to the transmitter. Multiple packets may be outstanding using a sliding-window protocol. If a packet is not acknowledged within a specific time, the transmitter backs up to the last successfully received packet and retransmits the following packets.

9. CALL INITIATION AND TRAINING

After call initiation, the pair of modems must exchange information and analyze line conditions to choose the optimum modulation methods, transmission speeds, codeword sets, spectral shaping parameters, and other features. All of this is done during a startup procedure involving several phases over a period of 10-30 s. Any modem user that has listened in on a data call will be familiar with the pings and chirps exchanged between the modems during these phases [13].

Phase	Description	Operations
1	V.8, V.8bis setup	Identify V.90 support, type of connection (digital or analog)
2	Probing	Exchange of modem capabilities; line probing; ranging
3	Half-duplex training	Equalizer and echo canceler training; digital impairment learning
4	Full-duplex training	Final training and fine-tuning.

The first phase allows the two modems to exchange information about their capabilities. It is during this phase that the modems will identify themselves as V.90-capable modems and whether they are in the role of the digital or analog modem. The protocols used in this phase are specified in ITU Recommendation V.8, which is also used by most other modem types, such as V.34. This allows full backward and forward compatibility between any combination of modem types.

Phase 2 is the probing phase and is identical to that used in V.34. During this phase signals are exchanged that allow the modems to provide details about their capabilities and to identify some characteristics of the transmission path, such as the round-trip delay, available bandwidth, and signal level. Also some of the parameters to be used in the subsequent training phases are chosen here.

The third phase is where most of the training of the equalizer and echo cancellers occurs and where the codeword set is chosen through a procedure known as *digital impairment learning*. By exchanging test signals the analog modem can determine the characteristics of the analog channel from the codec to the modem, lock its clock to the central office clock, and choose a subset, $\{C_i\}$, of the 256 possible codewords in each timeframe that make error-free decoding possible.

Phase 4, the final phase, allows exchange of the parameters chosen during prior phases such as the spectral shaping filter parameters, the elements of the codeword sets, and the overall transmission rates to be used. The parameters for the V.34 upstream transmissions are also provided during this phase.

In addition to the detailed interactions during each of these phases, the V.90 standard provides procedures for error recovery, retraining of the modems, and cleardown.

10. STANDARDIZATION

Standardization is an essential element of communications protocols. With adoptation of a standard, devices from multiple manufacturers can interoperate and achieve high levels of compatibility. For modems, the International Telecommunications Union (*www.itu.int*) in Geneva, Switzerland takes the lead role in overseeing this process. The name "V.90" refers to a Recommendation (the ITU uses this term rather than "Standard") from the ITU Telecommunication Standardization Sector (ITU-T). The recommendations in the "V" series all relate to data communication over the telephone network. Previous modem recommendations include the following:

- V.21 A 300-bps duplex modem standardized for use in the general switched telephone network
- V.22 A 1200-bps duplex modem standardized for use in the general switched telephone network and on point-to-point 2-wire leased telephone-type circuits
- V.32 A family of 2-wire, duplex modems operating at data signaling rates of up to 9600 bps for use on the general switched telephone network and on leased telephone-type circuits
- V.34 A modem operating at data signaling rates of up to 33,600 bps for use on the general switched telephone network and on leased point-to-point 2wire telephone-type circuits.

11. COMMERCIALIZATION

Modems have undergone an evolution not only in their algorithms and capabilities but also in the way they are implemented, packaged, and sold. Up until the mid-1990s, most modems were standalone devices that included all the hardware and software needed to connect between a telephone line and a serial port of a computer. In this "hard" configuration, the modem consists of interface circuitry, a controller, and a dedicated signal processor that runs the modem algorithms. Hard modems can be sold as external devices (box modems) that connect to a computer via a serial port, or as an add-on board such as a PCI card that interfaces directly to the computer's internal bus. In either case, hard modems do not require any data processing resources from the host computer.

To reduce costs, the host computer can instead handle some of the modem functions. For example, the controller, which provides the command set (usually the Hayes command set^4), can be moved from the modem to the host processor. The computer can process the commands and then direct the modem datapump operations, eliminating a microprocessor from the modem. The amount of processing power required to provide these functions is minimal, resulting in very little impact on the host computer. However, since this "controllerless" configuration requires a host computer to operate, it is sold as either an add-on board or as part of the computer itself, by integrating the modem onto the motherboard. The driver software for the modem contains the software needed to implement the controller, making the modem processor-dependent and operating-system-dependent.

The third generation of modems is the software modem. In this form, both the controller and the datapump operations are implemented on the host computer. This eliminates the signal processor from the modem reducing cost even further. The modem hardware then consists of only the DAA (the data access arrangement - an interface to the telephone line), and a bridge to the computer bus. All the signal processing is done on the host computer by executing software included with the driver for the modem. As with the controllerless modem, this type of modem is dependent on the processor type and the operating system. In addition, the datapump operations are much more CPU-intensive than the controller, resulting in reduced performance of the host computer for user operations while the modem is active. However, new protocols and enhancements can be added with a simple software upgrade of the driver.

By 2001, almost half of the 110 million V.90 modems shipped annually were soft modems with controllerless and hard modems sharing the remaining fraction [10].

12. PERFORMANCE

The V.90 standard supports downstream rates of \leq 56 kbps. In practice, however, these modems operate

⁴ The Hayes Command Set, which includes various commands prefixed by "AT," was introduced by Hayes Microcomputer for an early 300 baud modem and became a de facto industry standard. at lower speeds; 48–52 kbps is more typical. In part, this is due to the condition of the user's local loop. A long loop, or one that has anomalies, introduces distortions and noise that cannot be fully compensated for. The digital path through the telephone network may also introduce digital impairments that reduce the number of bits available for transmission. Another factor is the regulatory limits on the signal energy that reduce the available set of codewords such that speeds greater than 53 kbps are not possible even with very clean lines. During training, the modems choose a set of codewords and other parameters to maximize the data rate.

It is important to note that although data rate is used overwhelmingly as the critical performance measure, latency can have a greater effect on performance. In the typical use of a modem for an Internet connection, many protocols are layered on top of each other that require blocking of data or lookahead. Each of these layers adds delay in transmitting or receiving each byte of data. With TCP/IP, LAPM, spectral shaping, and other operations, round-trip delays between the endpoints can add up to more than 100 ms. For many separate accesses to small amounts of data, such as loading Webpages that contain embedded images, this latency, not the throughput, will dominate performance.

The actual rates and latencies obtained will depend on these factors as well as the details of the manufacturer's implementation. Although the standard fixes the protocol to permit interoperability, there is still great leeway in the implementations possible, creating significant performance differences between modem models.

13. FUTURE DEVELOPMENTS

ITU Recommendation V.92 [11] contains enhancements to V.90, most notably a significantly faster startup time. The negotiation and training time was reduced from as much as 45 seconds in V.90 to less than 15 s in V.92. This feature reduces not only the time a user has to wait for a connection but also the load on Internet providers by making it possible for users to drop and resume connections as needed. The faster reconnect times enabled another useful feature that V.92 adds: "modem-on-hold," which allows the modem call to be put on hold so the telephone can be used for a voice call. The data call can then be resumed where it was left off without dropping open connections.

Data transmission rates were improved in V.92 in two ways. Coding techniques that take advantage of the digital PCM network infrastructure have been applied to the upstream channel. Instead of using V.34 at rates of \leq 33.6 kbps, V.92 modems can transmit data upstream synchronously at rates of \leq 48 kbps while still allowing downstream rates of \leq 56 kbps. Improved compression based on the V.44 standard further improves data rates.

Since telephone calls are carried digitally within the telephone network at 64 kbps, it is not possible to go faster than this over a regular telephone line without modifications to the telephone network. Thus, there is unlikely to be any significant improvement in speed for telephone modems. However, it is possible to use the same physical wire on the local loop for much higher data rates if it is connected to different equipment at the central office. ISDN services provide two 64-kbps data channels and one 16-kbps signaling channel over the same twisted pair. Digital subscriber line (DSL) systems allow rates of ≤ 1.5 Mbps, and several systems that move data at rates exceeding 10 Mbps have been proposed—all using the same local loop. However, in all of these cases the local loop must be terminated at the central office with equipment different from that used currently for regular telephone lines.

Eventually, V.90 and V.92 modems will be replaced by DSL, cable modems, or other broadband access solutions for many data needs, such as home and business Internet access. However, the universality of basic telephone lines will ensure that telephone modems will remain present in laptops, settop boxes, games, and other devices for the foreseeable future.

BIOGRAPHY

Brent Townshend received a B.A.Sc. degree in engineering science from the University of Toronto in 1982. He received his M.S. in computer science, and M.S and Ph.D. degrees in Electrical Engineering all from Stanford University, Stanford, California, in 1983 through 1987. In 1987, he joined AT&T Bell Laboratories in Murray Hill, New Jersy, where he worked in the Acoustics Research Department on speech coding, speech recognition, and psychoacoustics. He also developed object-oriented programming systems and software architectures. Dr. Townshend moved to Montreal, Canada, in 1990 where he started Townshend Computer Tools. This company created new intellectual property through research and development, and then licensed the technology to other companies better suited to market the results. Example projects include the 56k modem, FAX coding systems, cryptography systems, and digital audio products. In 1997, Dr. Townshend cofounded Ordinate Corp in Menlo Park, California, a speech recognition/language testing company. Concurrently, Dr. Townshend has held adjunct or consulting professor positions at McGill University, Quebec, Canada, (1990-1994) and Stanford University (1994-2001) and is actively involved in the Silicon Valley venture capital community. Dr. Townshend has published numerous articles and over 20 patents. His current research interests include speech recognition, language assessment, music analysis, and digital photography.

BIBLIOGRAPHY

- C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27: 379-423, 623-656 (1948).
- D. V. Batorsky and M. E. Burke, 1980 Bell System noise survey of the loop plant, AT&T Bell Labs. Tech. J. 63(5): 775-818 (May-June 1984).
- I. Kalet, J. E. Mazo, and B. R. Saltzberg, The capacity of PCM voiceband channels, paper presented at IEEE Int. Conf. Communications, Geneva, 1993.
- 4. U.S. Patent 5,801,695, Sept., 1998, B. Townshend, High speed communications system for analog subscriber connections.

- 6. ITU Recommendation V.90, A Digital Modem and Analogue Modem Pair for Use on the Public Switched Telephone Network (PSTN) at Data Signaling Rates of up to 56,000 bit/s Downstream and up to 33,600 bit/s Upstream, International Telecommunication Union, Sept. 1998.
- T. Welch, A technique for high performance data compression, IEEE Comput. 8–19 (June 1984).
- ITU Recommendation V.42bis, Data Compression Procedures for Data Circuit-Terminating Equipment (DCE) Using Error Correction Procedures, International Telecommunication Union, Jan. 1990.
- 9. R. D. Gitlin, J. F. Hayes, and S. B. Weinstein, *Data Commu*nications Principles, Plenum Press, New York, 1992.
- Worldwide Modem Markets—a Semiconductor Perspective Year-to-date 2001 Market Wrap-up with Outlooks through 2005, VisionQuest, 2001.
- 11. ITU Recommendation V.92, Enhancements to Recommendation V.90, International Telecommunication Union, Nov. 2000.
- L. Brown and M. Davidson, PCM modem technology: Extending V.34, Commun. Syst. Design (Dec. 1997).
- L. Brown, PCM modem design: V.90 characteristics, Commun. Syst. Design (June 1998).
- ITU Recommendation V.42, Error-Correcting Procedures for DCEs Using Asynchronous-to-Synchronous Conversion, International Telecommunication Union, Oct. 1996.

VERY HIGH-SPEED DIGITAL SUBSCRIBER LINES (VDSLs)

J. CIOFFI Stanford University Stanford, California

1. INTRODUCTION

Few, if any, other technical inventions have had as much impact on mankind as Bell's invention of the telephone in 1876. Nearly a billion telephone lines exist worldwide, reliably enabling communication from nearly any point on earth with any other point. However, lesser known but potentially of yet greater and more lasting impact was Bell's invention of the less exciting twisted-pair cable in 1881. Since that time, twisted pairs have carried electrical phone signals reliably to those billion phones. But despite their relative age, twisted-pair lines have only begun to realize their potential to carry computer, television, radio, and other digital signals of the information age in addition to the plain old telephone service (POTS) they've quietly and reliably carried for over a century. Digital subscriber line (DSL) technology has emerged to service tens of millions of customers, with hundreds of millions envisioned before 2010. The latest in DSL technology is VDSL, which is overviewed in this article. Treatments of the earlier forms of DSL, such as the most heavily used and deployed ADSL, can be found in several books [1-5].

Very high-speed digital subscriber line (VDSL) service delivers very high bit rates over ordinary phone lines to customers. VDSL provides tens of megabits per second to those customers who desire broadband entertainment or data services while prudently leveraging infrastructure costs of fiber, avoiding wireless equipment placement, and rendering unnecessary coaxial-cable reengineering. VDSL modems can be programmed to carry symmetric (data network or LAN extension) or asymmetric (Internet Web surfing or TV) data rates over a variety of phone line types. Specific VDSL applications also encompass videoconferencing, telecommuting, telemedicine, distance learning, home shopping, and so on.

Figure 1 addresses the often asked question "How can DSLs go so much faster on the same phone lines that supposedly were already close to operating at theoretical limits with 33 kbps and 56 kbps voiceband modems?" Figure 1a illustrates the use of voiceband modems, specifically calling attention to the long transmission path that includes voiceband channels through telephone company switches. It is the bandwidth allocated by the switch to voice calls that limits the bandwidth of voiceband modems (i.e., the switch does not know it is a digital modem and treats the signal like voice). Of particular note, it is not the phone line that currently prevents transport of broadband data signals to the customer. Capacities of phone lines depend heavily on length, but a huge percentage are capable of carrying very high data rates if the narrowband switch can be avoided. DSLs (Fig. 1b), avoid the voice switch and instead have a transmission path that includes only the twisted pair. Digital signals are extracted by a second modem in the phone company central office (or a fiber-fed cabinet in the street) before they get to the switch and are then routed as broader bandwidth digital signals through an appropriate broadband network based on technologies such as ATM and IP. VDSL delivers the highest data rates on the shortest of the twisted pairs.

1.1. VDSL History and Basics

The VDSL concept was first published in 1991 [6], and was a result of a joint Bellcore–Stanford research study into the feasibility of 10+ Mbps symmetric and asymmetric data rates on short phone lines. The study specifically searched for the potential successors to the then more prevalent 1.5 Mbps HDSL and the then relatively new (then also only 1.5 Mbps) ADSL.¹ The first serious suggestions that VDSL be standardized first came almost simultaneously in the American T1E1.4 group from Amati Communications Corp. [7] and in the ETSI group from British Telecom [8] as a function of the first ADSL trials in Britain at 2 Mbps and 6 Mbps (supplied by Amati

¹ The author would like to gratefully acknowledge encouragement and support from then-retiring Dr. Joseph Lechleider of Bellcore, who encouraged and financially supported this early "VDSL" study.



Figure 1. DSL versus voiceband modems: (a) voiceband modem communication model; (b) DSL modem communication model.

to BT) where discussions on the potential of higher speeds on shorter fiber-fed ONU-based copper loops were active between the two companies. VDSL history also has connection to early high-speed ATM network studies in the ATM forum [9] and DAVIC [10], which attempted to transmit 26 and 52 Mbps symmetrically on one or more twisted pairs over very short distances (<100 ms) for localarea networks. While the latter ATM and DAVIC efforts are somewhat forgotten, in that instead 100base-T and now Gigabit Ethernet became the methods of ubiquitous use for internal computer networks on twisted pair, these early ATM and DAVIC efforts did also provide useful information to the development of present-day VDSL standards.

Asymmetric VDSL is viewed more as a residential service, introduced into the existing twisted-pair loop that carried only POTS or ISDN-BA (Integrated Services Digital Network, Basic rate Access) services. A general case of asymmetric service deployment is illustrated in Fig. 1b. Coexistence of POTS/ISDN and VDSL signals in the same twisted pair is allowed by the separation in frequency of their transmission bands, provided by the service splitter, shown as "split" in Fig. 1b.

Figure 2 presents the spectrum of the POTS/ISDN and VDSL signals running over the customer twisted pair. Fiber loop-carrier systems are being deployed worldwide to bring the high-bandwidth promise of fiber closer to groups of hundreds of phone company customers whom are served by what is called an "ONU" (Optical Network



Figure 2. Spectral allocation of VDSL and ordinary phone services.

Unit). Twisted pairs emanate from the ONU and connect to the customers, completing the path started with fiber to the ONU. Because of the service splitter operation, ordinary phone service appears the same to customers who seldom know that they are served by the ONU. (ONU's serve about 15% of the American population and often a smaller percentage in other countries.) The ONU is usually located less than about 1 km (3000 ft) from the customer. Desired asymmetric VDSL data rates over these distances are 13-26 Mbps downstream and 2-3 Mbps upstream, to allow delivery of digital TV (DTV) and high-definition TV (HDTV) services, superfast Web surfing and file transfer, and virtual offices at home. For shorter distances (\leq 300 ft) the downstream rate can be as high as 52-100 Mbps,

Symmetric VDSL is usually viewed as a business service, allowing 10-Mbps connections over a twisted pair of up to 1 (5000 ft) and 25 Mbps over shorter loops (<3000 ft). A typical application is Ethernet/IP LANs' interconnection with data rates of 100 Mbps (on 4 coordinated phone lines) or 10 Mbps on 1 or more pair, respectively, over a twisted pair between buildings in a corporate campus environment, as shown in Fig. 3 or between a telco central office and a business. Fiber may connect one central building with a network provider while the other buildings within the campus are connected by twisted pairs. Local area networks may service the individual buildings at 25.6 Mbps for ATM or 10 or 100 Mbps for Ethernet/IP over coax, fiber, wireless, twisted-pair, or other media. To expand the network between buildings making use of the existing phone lines, VDSL modems implement the high-speed connection. The traditional method to provide building-to-building connectivity requires use of many phone lines, each at 1.5 Mbps, with inverse multiplexers. This method, shown at the bottom of Fig. 3, is very expensive (inverse multiplexers are much more expensive than VDSL modems) and wastes twisted-pair bandwidth.

There has been an increased interest to use VDSL to transmit multiple T1 (1.5 Mbps), E1 (2 Mbps), and other T3 (45 Mbps) tributary datastreams to serve business customers in the area close to the local exchange (CO-central office). A typical deployment scenario, usually called "CO-based VDSL," is presented in Fig. 4. In this case, connections from the CO to the business are usually established on leased twisted pairs.

1.1.1. ADSL Extension. ADSL is now acknowledged as a successful telecommunications service with tens of millions of lines in deployment, and hundreds of millions to be deployed in the next decade or two. However, in its earliest days of standardization, ADSL faced the severe criticism that even its greatest standardized speed of 8 Mbps was too slow to match the data rates possible on what are called "hybrid fiber coax" (HFC) networks. HFC networks upgrade existing unidirectional cable-TV networks in two ways:

- 1. The cable TV networks are made two-way in HFC by replacing upstream-blocking filters in TV by more sophisticated two-way non-upstream-blocking filters.
- 2. The cable TV networks are increased in bandwidth in HFC by replacing coax near the TV head-end by



Figure 4. CO-based VDSL for business services.

fiber, multiplexing multiple separate coax signals on the same fiber in different bands, and thus rendering fewer subscribers per coaxial section (thus sharing less bandwidth).

Phone companies believed in 1994 and 1995 that they must replace their existing phone line networks with HFC, and several attempted to do so, only to later find costs prohibitive.

ADSL was already bidirectional, but with limited speeds downstream and yet more limited speeds upstream. (The asymmetry in ADSL allowing a longer line length for reliable transmission of a given data rate [1].) ADSL in 1994 and 1995 was perceived by telcos as too slow with respect to HFC. VDSL emerged from ADSL proponents as a next higher-data-rate step for ADSL—if fiber can be installed in HFC, then why not install it in existing networks when there are customers ready to pay for higher speeds than ADSL and instead use fiber-based loop shortening to increase the speed and symmetry of ADSL?

The initial VDSL architecture, shown in Fig. 5, ensued as the future of DSL deployment when (and if) customers were willing to pay for more and more fiber. The VDSL story is thus far more incremental in nature for telco deployment than HFC, which required an entire network to be replaced on the hope that enough customers would pay for it. In 1995 and beyond, this type of incremental DSL deployment won increasing favor with telephone service providers and is the actual mode of choice today. Cable suppliers continue to upgrade their TV networks to HFC at significant cost, but it becomes increasingly clear that the merits of DSL will prevail for nearly all services other than (unidirectional) analog and newer digital television delivery, for which cable seems to be still well conceived and currently the favorite.² The optional splitter of ADSL is preserved in VDSL so that analog voice service can be protected and preserved on the same line as VDSL. The cost of the fiber section is high, but can be divided by the number of customers served. As fiber penetrates closer and closer to the customer, that cost is shared by a smaller number of customers. Thus an important tradeoff in VDSL is the length of the fiber versus the length of the remaining copper. There is no single good answer to this tradeoff as it depends on applications, customer willingness to pay, transmission method, and, of course, the cost of the fiber-however, VDSL allows a wide range of trade-off as this chapter will illustrate

Figure 6 illustrates data rates for both upstream and downstream VDSL transmission on 24-gauge twisted pair (0.5 mm European) versus loop length for the American

 2 However, it is conceivable that Internet-based approaches to TV may allow an opportunity for DSL.



Figure 5. VDSL system architecture with ONU/fiber loop carrier system.



Figure 6. Downstream (**a**) and upstream (**b**) data rates for American and European standard DMT VDSL.

(998 curve) and European (997 curve) DMT VDSL standards.³ Clearly, the data rates are quite high on short loops, ensuring a greater individual bandwidth per user than cable networks (which customers must share inherently in architecture).

The premium paid in range loss for symmetry of data rate is less as loop lengths get shorter, and so then VDSL also offers a way to offer increasingly symmetric individual service to customers. As the number of small businesses worldwide explodes, most often in urban areas where line lengths are short, the potential for symmetric support of the voice, conferencing, peer-to-peer gaming or working, "home" Web server upstream bandwidths is then evident with VDSL. Today, an increasing number of service providers consider exploring early VDSL deployment for business services, particularly what is often called "fractional T3" support. In 2001, only 12,000 of the nearly 10,000,000 businesses in the United States were connected by fiber (and a smaller percentage in other countries). One large fiber installer and service provider estimates that they can increase this number by 2000 in the near future with cost of \$1 billion [17]. Thus, VDSL will play a major role in the future service offerings to small and many large businesses before fiber connection is financially viable

 $^{^3}$ Achievable data rates for the other "single-carrier" standard will be less, with these curves as an upper bound [1].

or completed. Other service providers still believe that support of video and television may be viable also in the future, although the economics of this application may be harder to justify versus cable.

The wealth of ADSL installations also then mandates another practical requirement that VDSL service must be compatible in many respects with existing ADSL. Existing customers with ADSL modems on their premises (perhaps in their portable computers) may move or travel into another area, or may live in an area where VDSL arrives, and will still want their ADSL modem to function as it always has. Thus, the ONU-side modem in Fig. 5, often called the LT (line termination) in VDSL, would need to support ADSL service, but would, of course, also allow higher speed service if/when that customer decides to purchase a higher-speed VDSL modem and the higher-speed VDSL service. Also, a customer who buys a VDSL modem will certainly want that modem to work with an existing ADSL connection at lower speed if that is all that is available. In addition to interoperation with ADSL modems, VDSL modems must also be compatible spectrally with ADSL modems that may share the same binder and with existing home premises networks, as well as with perhaps a plethora of other standard or nonstandard systems that exist in/near the cable (for instance, HDSL, SHDSL, or nearby ham radio). Subsection 7.1.3 deals more directly with this spectrum issue.

Overall, VDSL offers a mechanism for service providers to upgrade their networks incrementally and with continued profitability to include increasing amounts of fiber, approaching an ultimate goal of a network based entirely of fiber. Telephone line service providers have a very powerful story and future with VDSL, now following ADSL. The author suggests that perhaps the term "VDSL" will become synonymous with DSL in the near future.

2. VDSL ARCHITECTURE

With VDSL providing a significant growth opportunity for DSL in the future, the question becomes the details of "where, when, and how" to install fiber. Replacement of copper by fiber first occurs in cables closest to the central office. The cost of trenching for the new fiber in the cables closest to the central office can be shared by all the customers who are served by that cable of wires replaced. Further from the central office, newly installed fiber increasingly services fewer customers, ultimately just one customer when it connects to a specific customer premises. Thus, fiber installation cost increases per customer roughly with distance from the central office. In 2001, 15% of the American network had what is called "fiber feeder" as shown (and marked just "fiber") in Fig. 5. Virtually no fiber went to residential customers in 2001. Only about 1% of the many small businesses (about 10 million) in the United States are connected by fiber directly.⁴ Initially with the fiber loop carrier system being deployed today (Fig. 5), usually only the POTS/voice connections to residential and small business customers exist. However, VDSL or any DSL is increasingly added by placing the DSLAM⁵ at the end of the fiber, sometimes known as an optical network unit (ONU). Many phone companies have massive expenditures and fiber deployments under way to augment their ADSL service deployments so that line lengths are shortened and ensure higher ADSL speeds more reliably. The number of such DSL-enabled ONUs thus will grow rapidly in future years. This position of a DSLAM is often called a "line terminal" (LT), depending on the country and the exact type of DSL deployed (but unfortunately the use of these terms is inconsistent). We will use the term LT in the ensuing part of this article. VDSL-enabled DSLAMs in 2001 were just being introduced on a experimental basis in a few locations, but their number can be expected to increase especially as backward-ADSL-compatibile VDSL DSLAMs enter the market from major suppliers in 2002.

The trend toward more fiber and shorter twisted pairs thus will continue with time and VDSL. Splitter circuits can be used at both DSLAM and customer-premises ends of the line to preserve the existing POTS service in analog. Additionally, the high speeds of VDSL allow multiple digital voice signals to also be carried to the customer. Within the customer premises (home or business, home is shown in Fig. 5), a gateway is used to demultiplex the various VDSL signals and route them to the appropriate applications device, which could be a phone, computer, or television/entertainment system. Within the central office, another demultiplexer/multiplexor can be used to extract application signals and route them appropriately. Figure 5 presumes a heavy use of internet delivery, as well as Ethernet distribution within the home, but other mechanisms for such multiplexing are possible and discussed. In particular, wireless LANs [18] and/or homephone distribution systems [19] have also been used. The economic tradeoffs, speeds, demand for services in DSL make the actual tradeoff of length of fiber versus length of copper issue difficult to assess in 2001.

However, a point of service potentially is the so-called "distribution point" (or sometimes CSI), which typically is within 3000 ft of the customer, is shown in Fig. 4. This point is typically where larger cables are terminated and smaller cables servicing up to a few hundred customers begin. Usually, the box at the CSI basically serves as a cross-connection point for twisted pairs. However, the entire distribution-point box can be replaced if fiber feeds this CSI point. VDSL modems placed in such an enclosure then energize the subscriber-side twisted pairs that emanate. Power and size constraints are at least as difficult at this point as at the remote terminal, usually leaving a small area (a few square inches) and about 1 W of available power per DSL customer. Another intermediate point is yet closer to individual customers is often called the "cabinet" or "pedestal." Usually only 4-16 customers are served from the pedestal with individual twisted pairs

⁵ DSLAM is a *digital subscriber line access multiplexer*, a piece of equipment that houses several DSL modems at the service-provider end of the telephone line.

⁴ A large business typically has a small central office or other switching/routing mechanism within its largest campuses, and this switch will often be connected by fiber to the larger telecommunications network.

emanating to these customers. The pedestal again is normally a cross-connection point for telephone lines, but fiber can be deployed to this physically accessible point, and a VDSL modem deployed there. Very high speeds are possible on the resulting phone lines of 100 m or less, potentially hundreds of Mbps or more, higher yet than current VDSL.

Placing fiber to each successive point is increasingly costly because the cost of the fiber per subscriber necessarily increases as the number of customers decreases. Considerable "digging," "wall cracking," or physical labor may be necessary as the fiber proceeds closer to the customer. However, in the future if the customer demands higher bandwidth, then potentially higher revenue is possible also to pay for the fiber deployment costs. Ultimately fiber can be run to the home or even into the home to the desk/TV-top. The key to VDSL is the incremental deployment if and where customers are willing to pay for more fiber. The cost of deploying fiber can be from \$250,000 to \$1,000,000 per half-mile in areas of reasonable customer density.

2.1. Unbundling Issue

Colocation of VDSL modems is yet more difficult when the VDSL modems are not in a central office. This is because sharing of space by different service providers at the cabinet, carrier service area (CSA), or distribution point is physically difficult (there is not enough space). Today, this is a hotly debated issue in DSL deployment, and a single solution has not yet emanated. Some service providers accuse incumbent local exchange carriers of installing more fiber just to complicate colocation. Potential solutions for VDSL colocation are to

- 1. Standardize the backplane interface and card size(s) of VDSL so that many service providers may plug into an ONU.
- 2. Provide separate fibers to the ONU for each service provider and divide the existing small space according to the fibers that enter.
- 3. Use the HFT concept and colocate at the central office where more space is available.
- 4. Provide higher-level unbundling at layer 2 or 3 in the protocol stack.

Other solutions may evolve. VDSL standards to date have only encompassed colocation by mandating that a single spectrum type shall be used in all VDSL transmission types to minimize crosstalk between VDSLs. Largely, current VDSL standards are just beginning to address the intricacies of the VDSL colocation issue. In addition, the American ANSI group TIE1.4 has a new standards' effort in Dynamic Spectrum Management (DSM) that offers very attractive alternatives for all co-location alternatives.

2.2. POTS Splitters in VDSL

Splitter circuits for ADSL and VDSL are described in basic detail and design in Ref. 1, Chap. 3. For VDSL, the necessity of a splitter continues to receive attention. The heavy use of splitterless ADSL suggests that perhaps splitterless VDSL is also advisable for compatibility and volume deployment reasons. The first splitterless VDSL proposals appeared in Refs. 16 and 17—while these proposals in standards met with minority opposition (which is sufficient to block standardization), most advocates of the design in Section 3 are pursuing various forms of splitterless operation as an additional feature and option, albeit proprietary. The VDSL technology in Section 4 will not operate without splitters because of the consequent home wiring effects.

VDSL transmission designs on a splitterless channel will need to be robust to bridged taps, increasing amounts of radio interference, potential crosstalk (on same or other lines) from home services already present on the phone lines, and further signal attenuation. Such modems may also have need for control of power spectral density masks also to avoid excessive emission on the customer's premises that might interfere with local ham operators, emergency radio, or other appliances. The methods in Section 3 address these problems.

2.2.1. Common VDSL Reference Configurations. A common reference model was adopted and illustrated in Fig. 7 [20]. The interfaces and functionality associated with the two γ interfaces are common to both VDSL transmission methods. The PMS-TC (physical medium specific transmission convergence) and PMD (physicalmedium-dependent) are specified for each transmission method [21,22], while the spectra of the U interfaces and the splitter are again specified in common for the two transmission methods [20]. Like ADSL, VDSL also specifies two paths, a slow or interleaved path and a fast or noninterleaved path as in Figs. 8 and 9. The former undergoes interleaving as well as forward error correction to allow for maximum impulse noise protection while the latter allows for minimum delay (no more than 1 ms in VDSL). The application specific reference is the DSLAM device that basically makes use of a subset of the functionality for a given PMS-TC interface, converting from the γ interface. The γ can be an ATM or STM interface, for instance at speeds well above those of an individual DSL modem. The application specific layer then extracts the pertinent bits (presumably set up by the normal ATM method for setting permanent or virtual channels) for each of the fast and slow data paths through the VDSL modem, formats those bits into a known and reversible format within each stream and then forwards fast and slow bits to the PMS-TC interface.

Indeed, the greatest challenge of VDSL, presuming a working modem, may be the high-speed extraction and identification of the individual application signals, likely sent through an ATM or IP switching system. This section notes a few characteristics of interest for transmission.

One can envision the application devices as multiplexing and demultiplexing the applications signals, of which some may be simultaneously present in both the fast and slow buffers, and formatting them for/from the modem itself. The high-bandwidth data channel created by VDSL may allow for numerous applications to flow simultaneously.



Figure 7. VDSL reference model from wang [20].



Figure 8. VTU-O reference model.

The VDSL standard Part I [20] has an elaborate list of operational and maintenance capabilities to which we refer the reader. As with previous DSLs, the main parameters of interest are the state of the modems, the likelihood or presence of errors on the link, and the synchronization of network functions. VDSL allows passage of the 8-kHz network timing reference.

2.3. VDSL Spectrum Issues

As the highest speed DSL, VDSL uses the greatest amount of spectrum. Thus, it has the greatest concerns for spectrum compatibility. The issues of crosstalk and emissions from VDSL into surrounding telephone lines and radio receivers is more important and complicated in particular. Also, the crosstalk from existing services also affects VDSL spectrum design and performance. *Near-end crosstalk* (NEXT) is the radiation from one line's signals in one direction into to another line's signals moving in an opposite direction—in effect, a "near end" otherline's transmitter signals into the local receiver. Far-end crosstalk (FEXT) is the same effect but into signals going the same direction, thus a "far end's" other-line's transmitter signals into the local receiver. Furthermore, increasingly popular home LANS on twisted pair within customer premises will also complicate issues and tradeoffs, as there is both spectrum overlap on the same line with VDSL as well as crosstalk issues into other VDSL lines from the home-LANS. Considerable debate occurred in



Figure 9. VTU-R reference model.

standards meetings for the design of VDSL spectrum, and there are correspondingly three internationally approved spectrum plans (presumably one selected for any specific geographic region). Unfortunately, the competitive interests between incumbent and competitive service providers and the competitive interests between different transmission techniques did not work to VDSL's best spectrum advantage so far, as significant compromise can occur sometimes for nontechnical reasons or rationale. This area has been reopened several times as VDSL spectrum management standardization continues, and as advanced transmission enhancements alter basic parameters and issues. The three options do provide considerable flexibility for the future, although fortunately, as issues are revisited by technologists, marketing persons, and national regulators. This uncertainty makes this section at this time a bit difficult to write, but this chapter will focus on technical issues and describe the three plans, illustrating the various tradeoffs. In the long-term massive deployment of DSL, the service providers and equipment/chip vendors who best comprehend all the aspects of this area will be able to garner the best business advantages in massive DSL deployment. These spectral options appear in Fig. 10 and are discussed in the next subsection. The previously mentioned and very new DSM effort targets specifically these problems.

2.3.1. Spectral Plans. The need for a fixed spectrum plan is only necessary for compatibility of the "single-carrier" plans, Section 4 of this chapter, whereas the digital duplexing of the DMT spectrum allows arbitrary placement of band edges without excess-bandwidth penalty (although a 7.8% cyclic prefix penalty is necessary, see Section 7.3). It is possible that the plans in Fig. 10a,b will be replaced by those that fully consider all aspects of applications and deregulation in the future, or may be dynamic as in DSM. The 3rd international spectral

plan in Fig. 10c encompasses the possibility of spectrum flexibility. This option is implemented only in the DMT VDSL standard [22].

2.3.2. Robustness. VDSL must be able to accommodate frequency-selective disturbances, the best known of which are bridged taps of different lengths. Bridged taps occur in the loop plants of all operators (even though some try to deny it) and in particular are extremely pronounced in occurrence when splitterless designs are used. Immunity to bridged-taps is here called "robustness." Although it is impossible to avoid their effects completely, it is highly desirable for performance to degrade somewhat gracefully with bridged taps; for a symmetric service this can be interpreted as maintaining the ratio of upstream rate to downstream rate close to 1 even if total sum data rate up and down decreases slightly. In this symmetric case, huge rate loss in one of the directions because of bridged taps would be highly undesirable. In asymmetric transmission, it is desirable to maintain the ratio of asymmetry under different bridged-tap configurations.

First, this section illustrates the adverse effects of bridged-taps on transmission performance. Figure 11 shows the transfer function (in dB) of a 4050-ft loop, with bridged taps (66, 56, 46, and 36 ft long) and without bridged taps. The bridged taps cause the transfer function to exhibit notches periodically in frequency. As the bridged taps get shorter, the notches become more deep and move to higher frequencies. The existence of such notches (10-20 dB deep) can seriously harm transmission.

Below the graph, two different 4-band frequency plans are drawn. Note that this specific loop has very large attenuation at frequencies above 7 MHz, so the spectrum above 7 MHz is unsuitable for data transmission. Therefore, only the lower 2 bands would actually be used. Each frequency plan copes differently with this kind of



Figure 10. (a) plan 998—North American VDSL spectrum (U = upstream, D = downstream) (additional radio bands notched when used at 18.068-18.168, 21-21.45, 24.89-24.99); (b) plan 997—European VDSL spectrum (same unshown notched bands as 4a); (c) international flexible VDSL spectrum plan (f1, f2, f3, f4 determined programmably).

disturbance. If plan A were used in the presence of a bridged tap 36 to 66 ft long, then upstream transmission performance would be degraded significantly, although the downstream direction would not be affected. If plan B were used, then the downstream transmission performance would be degraded. Both plans fail to be robust.

One might argue that some other fixed 4-band plan would actually show more immunity to such situations. We explain why this is false: the bridged-tap length is not determined, so it may vary from 10 to more than 100 ft. This means that the notches may actually occur in almost any frequency of the VDSL spectrum. For any 4-band plan, there will always be a bridged tap with such a length that performance will solely be degraded in one direction. This direction is often the upstream direction using conventional models as are used in this article. However, there is one optimal frequency-division duplexing scheme, which one can prove attains the maximum possible robustness. The solution is to partition the spectrum into infinitesimally small bands and alternatively assign them to upstream and downstream transmission. Then, any frequency-selective disturbance (such as a bridged-tap) will have an equal impact on both directions of transmission. Figure 6 shows such a frequency plan, and illustrates why symmetric service is maintained.

Implementation of this optimal scheme may prove too complex,⁶ so suboptimal schemes with adequate robustness may have to be used instead. By interpolating between the 4-band plan and the optimal plan, we deduce that a number of bands as large as possible is highly desirable. As the number of bands increases, the data rate loss caused by a bridged tap will be distributed more evenly between the two directions of transmission. The simulations to follow demonstrate this fact.

2.3.2.1. Simulations. The simulation results that are shown below were obtained using a popular telco simulation tool. The 4 different frequency plans that were evaluated are shown below (numbers refer to MHz):

998 up = (3.75 - 5.2, 8.5 - 12)down = (0.138 - 3.75, 5.2 - 8.5)997 Up = (3.25 - 5.1, 7.1 - 12)Down = (.138 - 3.25, 5.1 - 7.1)Digital Duplexing 5-band plan up = (0.03 - 0.138, 3.08 - 4.78, 10.242 - 17.66) $down = complement \ of \ up$ Digital Duplexing 7-band plan up = (0.03 - 0.138, 2.5 - 3.5, 4.5 - 5.5, 11 - 17.66)down = complement of upDigital Duplexing 15-band plan up = (0.03 - 0.138, 2.1 - 2.5, 2.75 - 3, 3.25 - 3.5, 4 - 4.25, 4.5 - 4.75, 4.55-5.5, 10.5-17.66)down = complement of up

The services evaluated were medium symmetric, long symmetric, extralong symmetric, medium asymmetric, and long asymmetric in a noise A and noise D environment [18]. For each service the reach in meters was computed both with and without bridged taps. Table 1 shows

 $^{^6}$ Demonstrations of full zippering have been able to suggest that at least in some situations, large numbers of alternating up/down bands are indeed feasible with acceptable implementation.



the resulting reaches, and the percentage gains achieved by the plans using more than 4 bands in comparison to the 4-band plans.

We immediately see from Table 1 that using a larger number of bands always improves performance. A 4-band plan has upstream data rate annihilated with the 998 or 997 plans, a particularly concerning problem for those desired symmetric service, and for those who have been told that their wishes were accommodated by those plans. It is worth noting that the reach of the extralong symmetric service is improved by more than 35% (300–500 m), when more than 7 bands are used. This service represents an important market segment, and might be the first VDSL service to be deployed. A more detailed set of results appears elsewhere [19].

2.3.2.2. Mobile Radio Noise Robustness. Mobile radio noise robustness is important also to VDSL. The area

Figure 11. Illustration of robustness with bridged taps. The graph shows the insertion loss (in dB) of a 4050-ft loop with bridged-taps of length 66, 56, 46, and 36 ft (20, 17, 14, and 11 m, respectively). Below the graph, two different frequency plans are shown. When plan A is used, only upstream transmission is affected. When plan B is used, only downstream transmission is affected. In both cases symmetric service is disabled.

is sensitive because ingress transmissions can include those used in defense of various countries. Roughly, though, individuals without security clearances in various countries can learn of two types of disturbance:

- 1. Several narrowband analog voice signals in a band of a few hundred kilohertz that can be anywhere between 1 and 20 MHz and can move (or hop) in their general band location
- 2. Direct-sequence spread-spectrum signals of 300–500 kHz bandwidth (spread voice or data) with center frequency that hops throughout the 1–20-MHz band

One cannot specify in advance those bands to be annihilated by radio noise, and the joint operation of VDSL and of defense systems is highly preferred, especially in emergency situations.

Table 1. Bridged-Tap Robustness Results for 4500-ft 26-Gauge Loop

-					
ANSI Noise A	998	997	5-band	7-band	15-band
Rate 4500 ft	of 26 gauge				
Upstream Downstream	$\begin{array}{c} 40 \hspace{0.1 cm} \text{kbps} \\ 15.6 \hspace{0.1 cm} \text{Mbps} \end{array}$	$\begin{array}{c} 260 \hspace{0.1 cm} \text{kbps} \\ 15.4 \hspace{0.1 cm} \text{Mpbs} \end{array}$	1.69 Mbps 15.6 Mbps	2.68 Mbps 14.7 Mbps	3.37 Mbps 14.0 Mbps



Figure 12. Illustration of mobile radio noise robustness for two plans with 1.1-km loop, 20 VDSL FEXT, noises A and D, 0.4-mm line.

Again the solution is, as shown in Fig. 12, robust with VDSL duplexing. This section models a single data signal of width 500 kHz coupling into a phone line at a level sufficient to cause loss of use of the same band on that line. This level can vary from -80 to -110 dBm/Hz depending on the length of line and other system parameters. In this case, the duplexing plans were again evaluated. The loop simulated is again a 1.1 km, 0.4-mm loop, and noises A and D [20] were used in addition to the radio noise, along with 20 VDSL FEXT. The worst-case position of the radio noise actually was the same for both plans, 4-4.5 MHz. The loss is again less for the plan with more subbands.

The 7-band plan again robustly achieves over 7 Mbps symmetrically in all cases for noise A while the 4-band plan is only 1.4 Mbps. The relative drop in performance for the 4-band plan is also larger (even though absolute data rates are smaller because of analog duplexing). For noise D, the data rates are 2.5 Mbps for the universal band allocation plan and only 575 kbps for 4-band. Again the relative as well as absolute loss is larger for 4-bands because it is not robust to radio noise. A larger number of bands (more than 8) can actually increase the noise A result for the universal band allocation to over 8 Mbps symmetric, which may be of specific interest in Europe.

2.3.2.3. HPNA. Home Phone Network of America (HPNA) or in-progress standard draft G.pnt of the ITU [14] specify a use of the telephone line bandwidth between 5 and 10 MHz for internal home phone line networking. This spectrum of course overlaps HPNA, leading to signal

disruption of VDSL on the same line as well as generated large NEXT into neighboring VTU-Rs of other customers than the HPNA user.

Political maneuvering led to this issue being ignored by standards bodies where, for instance, the bands used by G.vdsl and G.pnt, standards documents produced by the same standards body, overlap. One reason for this ignorance was a stipulation by a few that unidirectional lowpass filters could be installed by every user of an HPNA network (presuming that the customer takes the time to locate his/her network interface somewhere outside his/her home and then properly installs the filter), even though that filter is not necessary for his/her internal computers to talk to one another. Remembering that all HPNA users would need to have such a filter self-installed before the very first VDSL could be installed at least begs the question as to the merit of such a solution.

A second, more elegant, solution, which does involve complexity increase for the VTU-R is described in Ref. 28 where G.pnt signals could be cancelled from a VDSL signal when on the same line or on neighboring lines as long as there were only 1 or 2 of significant amplitude. G.pnt systems however do not appear to be gaining true market acceptance, and so this may be less of a problem for VDSL. If they do, interference cancellation of G.pnt by VDSL may become a necessity in practice.

Note that this spectral incompatibility concern does not apply to Ethernet, which is typically installed on category 5 wiring that is separate and isolated by nature and design from the telephone company network. Thus, even though Ethernet uses the same band, there is no actual spectral overlap on physically colocated wires.

2.4. The Grand Debate

Dating to the days of ADSL standardization, there has been a debate over the best transmission technology to use for DSL. While ADSL standards have universally selected the specific multicarrier transmission method known as DMT after considerable deliberation and testing, and a universal consensus, a few CAP (carrierless amplitude and phase) and QAM proponents nevertheless marketed nonstandard ADSL modems for a significant time period, before most switched to and supported standardized DMT. Subsequent debate was heated in the marketplace, and there were several attempts to reverse ADSL standards (from DMT to CAP/QAM) that were abortive. The supposed threat of fundamentally high complexity of DMT leading to high prices eventually was unequivocally refuted in the ADSL marketplace, where low-cost components abound today. Where VDSL was originally intended as an extension to the ADSL DMT standard [7], VDSL then instead became another chance for standardization of the QAM/CAP technologies in DSL. Because the American T1E1.4 group by-laws prohibited standardization in 1994 on line rates above 10 Mbps, that group's charter had to be rewritten for a new limit of 100 Mbps. As the charter was rewritten, CAP proponents insisted that the VDSL area be a new standard and that the line code issue be revised, even though the original VDSL was intended to extend ADSL speed and symmetry. Thus, the opportunity for vet another debate unfortunately emerged and continues to date.

This newer debate has continued in VDSL for many years with two large industrial consortia emerging with two complete transmission specifications for VDSL:

- 1. VDSL Alliance discrete multitone transmission (DMT) with digital or analog duplexing
- 2. *VDSL Coalition* "single"-carrier modulation(SCM) with analog duplexing

("Single" is in quotes here because the VDSL Coalition actually advocates a solution with two carriers in each direction, and an optional third carrier upstream.) Both groups have contributed a "temporary working" standard to T1E1.4, which appear in three documents: a common reference document [20], an SCM document [21], and a DMT document [22]. The DMT specification supports up to 4096 4.3125-kHz ADSL-style tones, and any number of up/down stream transmission bands. The lower 256 tones are exactly the same as ADSL, facilitating backward compatibility. Both groups have about 50 companies in them, with about 10 common members. The companies in both groups represent an enormous cross-section of the telecommunications world.

The VDSL Alliance solution is backward-interoperable with ADSL and has taken greater time to develop into its present converged state [22], but offers some outstanding flexibility and performance features in a large number of possible configurations. Some vendors sell chips that implement up to 32 ADSL modems or up to 4 VDSL modems. The VDSL Coalition specification [16] is slightly simpler to understand (although more pages in reality) and to design to, but not interoperable with ADSL. The Coalition specification had the advantage of earlier availability of transmission components that were partially compliant with it. International standardization in the ITU has steadfastly held to the principle that only one will become an international standard. The T1E1.4 group will revisit the two standards for permanency in 2003, when it is likely only one will survive. The IEEE 802.3 group on Ethernet in the First Mile (see Section 5) currently is also considering the two VDSL standards for that decision. At time of writing, it appears that the considerable advantages of backward interoperability with ADSL (now installed in over 20 million locations) and the spectral flexibility of the DMT approach will again cause it to prevail over the single-carrier method, and the number of single-carrier supporters dwindle to two or three companies. Clearly DMT is the best solution technically and from many other perspectives. However, the politics of standards groups can sometimes lead to tragic decisions, and this issue is not yet formally decided.

3. DMT PHYSICAL-LAYER STANDARD

Discrete multitone (DMT) transmission is the only worldwide standard for ADSL and VDSL transmission. DMT is the most efficient of the high-performance transmission methods that allow a transmission system to perform near the fundamental limit known as capacity [1]. For difficult transmission lines, there is no other costeffective high-performance alternative presently, and VDSL has the most difficult transmission environment of all DSLs. Variants of DMT for wireless transmission (known as OFDM) have also come into strong use in the area of wireless local area networks (IEEE 802.11(a), [13]) and wireless broadband access (IEEE 802.16 [25]), as well as digital terrestrial television (HDTV and digital TV) broadcast [26] in most of the world, all of which are known to also be particularly difficult transmission problems. The standardized VDSL DMT method is a natural extension of the method used for ADSL and backward compatible with it, as described in Section 3.1. Section 3.2 describes the "zipper" duplexing method that is also called "digital duplexing," which is an enhancement to the original DMT method that allows the upstream and downstream transmissions to be compactly placed in the limited transmission bandwidth of a telephone line. Section 3.4 investigates initialization.

3.1. Basic Multicarrier Concept

Starr et al. explained the basic multicarrier transmission concept of dividing a transmission band into a large number of subcarriers and adaptively allocating fractions of total energy and data rate to each to match an individual line characteristic [1]. It is this adaptive loading feature that sets multicarrier methods in a higher league of performance than other DSL transmission methods. VDSL presents a highly variable transmission environment with bandwidths that can vary from a few MHz to nearly 20 MHz, with intervening radio interference in several narrow bands, with huge spectrum notching effects from bridged taps, and with a variety of crosstalking situations. VDSL is undoubtedly the most difficult and highly variable transmission problem yet faced by DSL engineers. Any thing less than an excellent design will risk the viability of the DSL industry in the future, and multicarrier methods meet that challenge.

The first challenge for VDSL is interoperability with existing ADSL. One of the applications for VDSL is simply speed extension of ADSL, meaning that it is possible for an existing ADSL customer to have that service provided by a new ONU in his/her neighborhood that is VDSLready. A VDSL modem in the ONU that will interoperate with that existing ADSL modem is highly desirable so that no extra labor or purchases are necessary at the customer's premises should that customer elect to rest with their current ADSL service for a period of time before then electing to move to VDSL to increase the speeds of their ADSL service. Similarly, an existing ADSL customer may elect to purchase a VDSL modem (or may have one from a previous residence or business address) and then needs to interoperate with an existing ADSL CO modem. Thus, a requirement for incremental DSL rollout to higher speeds and increasing use of fiber is that VDSL interoperate with existing ADSL, meaning the lower 256down/32-up DMT tones of an ADSL modem must also be implemented by an interoperating VDSL modem. For this reason, the DMT VDSL standard [22] uses the same tone spacing of 4.3125 kHz that was used in ADSL. The VDSL standard allows for the DMT VDSL modem to symmetrically use numbers of tones of 256, 512, 1024, 2048, and 4096 or 2^{n+8} n = 0, 1, 2, 3, 4. The number 2048 is considered a default for full compliance with other VDSL modems, but interoperation with smaller numbers of tones is illustrated in Fig. 13, where it is clear whether upstream or downstream, the modem with the smallest number of DMT tones then dictates the maximum that can be used in that direction. Such elimination of superfluous tones can occur naturally during training, or may be selectively programmed during special initialization exchanges; the latter is usually the preferred implementation.

The following terms apply here:

Extension. The cyclic extension [1] size for DMT VDSL is optionally programmable to sizes $m \times 2^{n+1}$, where *m* is an integer. The modem must be able

to implement at least the default of $20 \times 2^{n+1} = 40 \times 2^n$ for the case of n = 0, which corresponds to interoperability with ADSL.⁷ Longer cyclic prefixes on shorter channels allow some additional performance-enhancing features to be added for DMT VDSL that are not implemented in ADSL, which are described in more detail in Section 3.2.

- *Encoder*. The constellation encoder for DMT VDSL and tone-ordering procedures are identical to those of the worldwide ADSL standard [1], although implemented perhaps over a larger set of tones for VDSL.
- *Pilot*. The pilot of ADSL has been made optional and generalized in VDSL. In ADSL, the pilot was sent downstream always on tone number 64 (276 kHz). In VDSL, the VTU-R can decide to use a pilot on any (or no) tone in initialization. If a pilot is used, the 00 point in the standardized 4point QAM constellation on that selected tone is sent in all symbols. The synchronization symbol of ADSL has been eliminated in VDSL, except when interoperating with an older ADSL modem.
- *Timing Advance.* The VTU-R is capable of changing the symbol boundary of the downstream DMT symbol it receives by a programmable amount, which is communicated during initialization to the VTU-O modem. This is also a new feature for VDSL that is used for implementation of digital duplexing as described in Section 3.2.
- Power Backoff (PBO). PBO has been studied by standards groups as a way to prevent upstream FEXT from a customer closer to an ONU from acting as a large noise for a VDSL customer further away. The basic problem is that the closer user could operate at a higher data rate or with better performance than is necessary or fair to other customers. Various methods for reducing the disparity among lines vary from introducing a flat power backoff at all frequencies as a function of measured received signal [21] to

⁷ Actually, in this case, the cyclic prefix is reduced by 8 samples if the synchronization symbol of ADSL is to be inserted every 69 symbols or 17 ms.



Figure 13. Interoperability diagram for standardized DMT modems of increasing speeds.

spectrally shaped methods that attempt to apply either the (1) reference noise method — forcing all upstream transmissions to have a FEXT of common spectral shape (and thus harm), or the (2) reference length method — forcing all upstream transmissions to have a FEXT the same as that of a nominal "reference" length VDSL line. The latter two methods, particularly (2), seem to have won favor with standards groups, but the area is still debated at time of writing. A method making use of actual line measurements [40] was introduced for future spectrum management and essentially eliminates the PBO issue, but came after VDSL standards had entered final voting and could not yet then be standardized.

Express Swapping. The standardized bit swapping of ADSL is mandatory also in VDSL [1]. However, VDSL also offers a highly robust and high-speed optional capability of instead altering the bit distribution all at once (instead of one tone at a time as in the older bit swapping). This is known as express swapping. The additional commands are described in the VDSL standards documents [22]. However, express swapping allows the new bit table is sent in one command and protected by CRC check-if correctly received, all tones are replaced with the new bit distribution at an immediately succeeding point specified in the commands and protocol. This allows a system to react very quickly to abrupt transients caused by excitation of crosstalkers or RF interferers, or perhaps an offhook line change in splitterless operation. It also enables advanced spectrum management features that may occur in the future.

3.2. Digital Duplexing

This subsection describes *digital duplexing*, which is a method for minimizing bandwidth loss in separating downstream and upstream DMT VDSL transmissions. Originally, this method was introduced by Isaksson, Sjoberg, Nilsson, Mesdagh, and others in a series of papers that refer to the method as "zipper" [27,34-37]. General principles, as well as some simple examples, illustrate how digital duplexing works and why it saves precious bandwidth in VDSL. This description is intended for readers familiar with basic discrete multitone transmission (DMT). Thus, readers can use their DMT knowledge and the examples and explanations of this section to understand the relationship of the cyclic suffix to the cyclic prefix, and thus consequently to comprehend the benefits of symbol-rate loop timing and to appreciate the use of windows without intermodulation loss.

Excess bandwidth is a term used to quantify the additional dimensionality necessary to implement a practical transmission system. The excess-bandwidth concept is well understood in the theory of intersymbol interference where various transmit pulseshapes are indexed by their percentage excess bandwidth. In standardized and implemented DMT designs for ADSL, for instance, the symbol rate is 4000 Hz while the tone width is 4312.5 Hz, rendering the excess bandwidth

(0.3125/4) = 7.8%. In ADSL, additional bandwidth is lost in the transition band between upstream and downstream signals when these signals are frequencydivision-multiplexed. In VDSL, this additional bandwidth loss is zeroed through an innovation [27] known here as *digital duplexing*, which particularly involves the use of a "cyclic suffix" in addition to the well-known "cyclic prefix" of standardized DMT ADSL. This subsection begins with a review of basic DMT and of its extension with the use of the cyclic suffix, including a numerical example that illustrates symbol-rate loop timing. This discussion illustrates why the time-domain overhead is all that is necessary to allow full use of the entire bandwidth without frequency guard bands in a very attractive and practical implementation.

This section proceeds to investigate crosstalk issues both when other VDSL lines are synchronized and not synchronized. Windowing and its use to mitigate crosstalk into other DSLs or G.pnt are also discussed, as are conversion device requirements.

This section then continues specifically to use a second example that compares a proposed analog duplexing plan for VDSL with the use of digital duplexing in a second proposal. In particular, 4.5 MHz of excess bandwidth is necessary in the analog duplexing while only the equivalent of 1.3 MHz is necessary with the more advanced digital duplexing. The difference in bandwidth loss of 3 MHz accounts for at least a 6–12-Mbps total data rate advantage for digital duplexing in the example, which provides a realistic illustration of the merit of digital duplexing.

3.2.1. Basic DMT. Figure 14 illustrates a basic DMT system for the case of baseband transmission in DSL [7], showing a transmitter, a receiver, and a channel with impulse response characterized by a phase delay Δ and a response length ν in sample periods. N/2 tones are modulated by QAM-like two-dimensional input symbols (with appropriate *N*-tone conjugate symmetry in frequency) so that an *N*-point inverse fast fourier transform (IFFT) produces a corresponding real baseband time-domain output signal of *N* real samples.

For basic DMT, the last L = v of these samples are repeated at the beginning of the packet of transmitted samples so that N + L samples are transmitted, leading to a time-domain loss of transmission time that is L/N. This ratio is the excess bandwidth. The minimum size for L = vis the channel impulse response duration (in sampling periods) for basic DMT (later for digitally duplexed DMT, L will be the total length for all prefix/suffix extensions and thus greater than v). Sometimes DMT systems use receiver equalizers [1] to reduce the channel impulse response length and thus decrease v. Such equalizers are common in ADSL. The objective is to have small excess bandwidth by decreasing the ratio v/N. In both DMT ADSL and VDSL, the excess bandwidth is 7.8%.

The IFFT of the DMT transmitter implements the equation

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n \cdot e^{j(2\pi/N)nk}$$



Figure 14. Basic DMT transmission system.

where x_k k = 0, ..., N - 1 are the N successive timedomain transmitter outputs (the prefix is trivial repeat of last v). The values X_n are the two-dimensional modulated inputs that are derived from standard QAM constellations (with the number of bits carried on each "tone" adaptively determined by loading [6] and stored in the b_n , g_n tables at both ends). To produce a real time-domain output, $X_n = X_{N-n}^*$ in the ubiquitous case that N is an even number. In ADSL, N = 512, while in VDSL, $N = 512 \cdot 2^{n+1}$ n = 0, 1, 2, 3, 4. One packet of $N + \nu$ samples is transmitted every T seconds for a sampling rate of N + L/T. With 1/T = 4000 Hz in ADSL and VDSL, the sampling rates are 2.208 MHz and up to 35.328 MHz, respectively, leading to a cyclic prefix length of v = L = 40samples in ADSL and a cyclic extension length of up to L = 640 samples in VDSL. The extension length in VDSL also includes the cyclic suffix to be later addressed.

When the cyclic extension length is equal to or greater than the impulse response length of the channel $(L \ge \nu)$, DMT decomposes the transmission path into a maximum of N/2 independent simple transmission channels that are free of intersymbol interference and that can be easily decoded. The receiver in Fig. 14 extracts the last N of the N+L samples in each packet at the receiver (when no cyclic suffix is used) and forms the FFT according to the formula

$$Y_n=\sum_{k=0}^{N-1}y_k\cdot e^{-j(2\pi/N)nk}$$

where the reindexing of time is tacit and really means samples corresponding to times $k = L + 1, \ldots, L + N$ in the receiver. The receiver must know the symbol alignment and cannot execute the FFT at any arbitrary phase of the symbol clock. If the receiver were to be somehow offset in timing phase, then time-domain samples from another adjacent packet would enter the FFT input, displacing some corresponding time-domain samples from the current packet. For instance, suppose the FFT executed m sample times too late, then the output would be

$$ilde{Y}_n = \sum_{k=0}^{N-m-1} y_{k+m} \cdot e^{-j(2\pi/N)nk} + \sum_{k=N-m}^{N-1} u_{N-m+k} \cdot e^{-j(2\pi/N)nk}$$

where u_k are samples from the next packet that are unwanted and act as a disturbance to this packet. Furthermore, *m* samples from the current packet were lost (and the rest offset in phase). Thus

$$\tilde{Y}_n = Y_n + E_n$$

where E_n is a distortion term that includes the combined effects of u_k , the missing terms y_k , and the timing offset in the packet boundary. $E_n = 0$ only (in general) when the correct symbol alignment is used by the receiver FFT. DMT systems easily ensure proper phase alignment through the insertion of various training and synchronization patterns that allow extraction of correct symbol boundary.

It is important to note that the FFT of any other signal with the same N might also have such distortion unless the symbol boundaries of that signal and the DMT signal were coincident. In the later case of time coincidence, the FFT output is simply the sum of the two signals' independent FFTs. Indeed the receiver would have no way of distinguishing the two signals and would simply see them as the sum in the time-coincident case.

The second signal could be the opposite-direction signal leaking through the imperfect hybrid in VDSL. If the transmitted and received symbols are aligned in time and frequency-division duplexing is used, tones are zeroed in one direction if used in the other. Then, the sum at the FFT output is simply either the upstream or the downstream signal, depending on the duplexing choice for the set of

indices n. No zeroed tones are necessary between upstream and downstream frequency bands as that is simply a waste of good undistorted DMT bandwidth. No analog filtering is necessary-the IFFT, cyclic prefix, and the FFT do all the work if the system is fortunate enough to have time coincidence of the two DMT signals traveling in opposite directions. The establishment of time coincidence of the symbols at both ends of the loop is the job of the cyclic suffix, which the next subsection addresses. In other words, the FFT works on any DMT signal of packet size N samples, regardless of source or direction as long as the packet is correctly positioned in time with respect to the FFT. This separation is not easily possible unless the DMT signals are aligned-thus VDSL DMT systems ensure this alignment through a cyclic suffix to be subsequently described.

Table 2 provides a comparison and summary of DMT use in ADSL and in VDSL. Note that DMT VDSL uses digital frequency-division multiplexing (FDM) and spans at most 16 times the bandwidth at its highest bandwidth use. This full bandwidth form is actually optional and the default values are shown in parentheses on the right, with the default actually being exactly half full.

3.2.2. Cyclic Suffix. The cyclic suffix occurs at the end of a DMT symbol (the opposite side of the cyclic prefix) and could repeat, for instance, the first 2Δ samples of the DMT symbol (not counting the prefix samples) at the end as in Fig. 16, where Δ is the phase delay in the channel (phase delay or absolute delay, not group delay, which is related to v). A symbol is then of length N + L. The value for L must be sufficiently large that it is possible to align the DMT symbols, transmit and receive, at both ends of the transmission line. Clearly alignment at one end of a loop-timed line⁸ is relatively easy in that the line terminal (LT) for instance need wait only until an upstream DMT symbol has been received before transmitting its downstream DMT symbols in alignment on subsequent boundaries. Such single-end alignment is often used by some designers to simplify various portions of an ADSL implementation. The alignment is not necessary unless digital duplexing is used. However, alignment at one end almost surely forces

⁸ Loop timing of the sample clock means that the NT (VTU-R) uses the derived sample clock from the downstream signal as a source for the upstream sample clock in DMT.

misalignment of symbol boundary at the VTU-R as in Fig. 15.

In Fig. 15, let us suppose that N = 10, $\nu = 2$, and that the channel phase delay or overall delay is $\Delta = 3$. The reader can pretend they have a master time clock and that the LT begins transmitting a prefixed DMT symbol at time sample 0 of that master clock. The first two samples at times 0 and 1 are the cyclic prefix samples, followed by 10 samples of the DMT symbol, that is, at times 0-11. At the receiver, all samples undergo an absolute delay of 3 samples (in addition to the dispersion of v = 2 samples about that average delay of 3). Thus, the cyclic prefix' first sample appears in the VTU-R at time 3 of the master clock and the DMT symbol then exists from time 3 to time 14. The samples used by the receiver for the FFT are samples 5-14, while samples 3 and 4 are discarded because they also contain remnants from a previous DMT symbol. The LT has DMT symbol alignment in Fig. 13, so that it also received the upstream prefix first sample at time 0 and continued to receive the corresponding samples of the upstream DMT symbol until time 11. Thus, the DMT symbols are aligned at the LT. However, in order for the upstream DMT symbol to arrive at this time, it had to begin at time -3 in the VTR-R. Thus the upstream symbol transmitted by the VTU-R occurs in the VTU-R at times -3 through 8 of the reader's master clock. Clearly, the DMT symbols are not aligned at the VTU-R.

In Fig. 16, a cyclic suffix of 6 samples is now appended to DMT symbols in both directions, making the total symbol length now 18 samples in duration (thus slowing the symbol rate or using excess bandwidth indirectly). The LT remains aligned in both directions and transmits the downstream DMT symbol from master clock times 0-17. These samples arrive at the NT at times 3-20of the master clock, and valid times for the receiver FFT are now 5-14, 6-15, ..., 11-20. Each of these windows of 10 successive receiver points carry the same information from the transmitter and differ at the FFT output by a trivial phase rotation on each tone that can easily be removed. The upstream symbol now transmits corresponding valid DMT symbols from time -1 to time 8, also 0-9, 1-10, and the last valid upstream symbol is time 5-14. At the VTU-R, the first downstream valid symbol boundary from samples 5-14 and the last upstream valid symbol boundary, also at samples 5-14, are coincident in time-thus the receiver's FFT can correctly find both LT and VTU-R transmit signals by executing at sample times 5–14 without distortion of or interference between tones.

Table 2. Comparison of DMT for ADSL and VDSL

	ADSL	VDSL (default)
FFT size	512	8192 (4096)
# of tones	256	4096 (2048)
Cyclic extension length	L = v = 40	$L = 640 \ge v + \Delta$
Sampling rate	2.208 MHz	35.328 (17.664) MHz
Bandwidth	1.104 MHz	17.664 (8.832) MHz
Duplexing	Analog FDM	Digital FDM
Tone width	4.3125 kHz	4.3125 kHz
Excess bandwidth	$86 \ kHz \ (prefix) + 40 \ kHz \ (filters)$	$1.3\ MHz\ (650\ kHz)\ (no\ filters)$



Figure 15. DMT symbol alignment without cyclic suffix.

At this time only (for cyclic suffix length 6), the receiver FFT and IFFT can be executed in perfect alignment, and thus downstream and upstream signals are perfectly separated. This DMT system in Fig. 16 has *symbol-rate loop timing*, and thus a single FFT can be used at each end to extract both upstream and downstream signals without distortion. (There is still an IFFT also present for the opposite-direction transmitter). This loop is now digitally duplexed.

Note that the cyclic suffix has length double the channel delay (or equivalently was equal to the round-trip delay) in the example. More generally, one can see that from Fig. 16 that VTU-R symbol alignment will occur when the equivalent of time 5, which is generally $\nu + \Delta$, is equal to the time $-3 + 2 + L_{suf}$, $= -\Delta + \nu + L_{suf}$. Equivalently $L_{suf} = 2\Delta$. In fact, any $L_{suf} \ge 2\Delta$ is sufficient with those cyclic suffix lengths that exceed 2Δ just allowing more valid choices for the FFT boundary in the VTU-R. For instance, the designer had chosen a cyclic suffix length of 7 in our example, then valid receiver FFT intervals would have been both 5–14 and 6–15. This condition can be halved using the timing advance method in the next subsection.

3.2.3. Timing Advance at LT. Figure 17 shows a method to reduce the required cyclic suffix length by one-half to $L_{\text{suf}} \geq \Delta$. This method uses a *timing advance* in the LT modem where downstream DMT symbols are advanced by Δ samples. The two symbols now align at both ends at time samples 2-11, and the cyclic suffix length is reduced to 3 in the specific example of Fig. 17. Thus, for VDSL with timing advance, the total length of channel impulse response length and phase delay must be less than $18 \,\mu s$, which is easily achieved with significant extra samples for the suffix in practice. In VDSL, the proposal for Lis 640 when N = 8192. The delays of even severe VDSL channels are almost always such that the phase delay Δ plus the channel impulse response length ν are much less than the 18-µs cyclic extension length (if not, then a time equalizer [6]). One notes that the use of the timing advance causes the transmitters and receivers at both ends to all be operational at the same phase in the absolute time measured by the master clock.

Digital duplexing thus achieves complete isolation of downstream and upstream transmission with no frequency guard band—there is however, the 7.8% cyclic extension penalty (which is equivalent to 1.3 MHz loss


Figure 16. DMT symbol alignment at both LT and NT through use of suffix.



Figure 17. Illustration of cyclic suffix and LT timing advance.

2798 VERY HIGH-SPEED DIGITAL SUBSCRIBER LINES (VDSLs)

in bandwidth in full VDSL and 650 kHz loss in the default or "lite" VDSL). Thus it is not correct, nor appropriate, to place frequency guard bands in studies of DMT performance in VDSL.

Digital duplexing in concept allows arbitrary assignment of upstream and downstream DMT tones, which with FDM VDSL means that these two sets of upstream and downstream tones are mutually exclusive. On the same line, there is no interference or analog filtering necessary to separate the signals, again because the cyclic suffix (for which the equivalent of 1.3 MHz of bandwidth has been paid or 650 kHz in default VDSL) allows full separation even between adjacent tones in opposite directions via the receiver's FFT. However, while theoretically optimum, one need not "zipper" the spectrum in extremely narrow bands, and instead upstream and downstream bands consisting of many tones may be assigned as described earlier. Some alternation between up and down frequencies is universally agreed as necessary for reasons of spectrum management and robustness, although groups differ on the number of such alternations.

3.2.4. Crosstalk. Analysis of NEXT between neighboring VDSL circuits needs to consider two possibilities:

- 1. Synchronization of VDSL lines
- 2. Asynchronous VDSL lines

The first case of synchronous crosstalk is trivial to analyze and implement with FDM. Adjacent lines have exactly the same sampling clock frequency (but not necessarily the same symbol boundaries). There is no NEXT from other synchronized VDSLs with FDM, as will be clear shortly. The second case of asynchronous VDSL NEXT (from other VDSL lines) into VDSL is more interesting. In this case, the sidelobes of the modulation pulseshapes for each tone are of interest.

A single DMT tone consists of the sinusoidal component

$$\begin{split} x_n(t) &= \left[X_n \cdot \exp\left(j\frac{2\pi}{N} \cdot \frac{N+\nu}{T} \cdot nt\right) \right. \\ &+ X_{-n} \cdot \exp\left(-j\frac{2\pi}{N} \cdot \frac{N+\nu}{T} \cdot nt\right) \right] \cdot w_T(t) \\ &= 2|X_n| \cdot \cos(2\pi f_0 nt + \angle X_n) \cdot w_T(t) \end{split}$$

where $f_0 = 2\pi/N \cdot (N + v)/T$ or 4.3125 kHz in ADSL and VDSL, and $w_T(t)$ is a windowing function that is a rectangular window in ADSL, but is more sophisticated and exploits digital-duplexing's extra cyclic suffix and extra cyclic prefix in VDSL. Figure 18 shows the relative spectrum of a single tone with respect to an adjacent tone for the rectangular window.

Note the notches in the crosstalker's spectrum at the DMT frequencies, all integer multiples of f_0 . Thus, the contribution of other VDSL NEXT will clearly be zero if all systems use the same clock for sampling, regardless of DMT symbol phase with respect to that clock. This is an inherent advantage of DMT systems with respect to themselves since it is entirely feasible that VDSL modems in the same ONU binder group could share the same clock and thus have no NEXT into one another at all. Indeed, this is a recommended option in [22].

When the sampling clocks are different, however, the more that sampling clocks of VDSL systems differ, the greater the deviation in frequency in Fig. 18 from the nulls, allowing for a possibility of some NEXT.



Figure 18. Magnitude of windowed sinusoid versus frequency; note notches at DMT frequencies.

Studies of such NEXT for DMT digital duplexing are highly subjective and depend on assumptions of clock accuracy, number of crosstalkers with worst-case clock deviation, and the individual contribution to NEXT transfer function of each of these corresponding worst-case crosstalkers. Nonetheless, reasonable implementation renders NEXT of little consequence between DMT systems.

If the VDSL PSD transmission level is S = -60 dBm/Hz, and the crosstalk coupling is approximated by $(m/49)^{.6} \cdot 10^{-13} \cdot f^{1.5}$ for *m* crosstalkers the crosstalk PSD level is

$$S_{\text{xtalk}}(f) = S \cdot (1/49)^{.6} \cdot f^{1.5} \cdot \left| \sin c \left(\frac{f}{f_0} \right) \right|^2$$

The peak or sidelobes can be only 12 dB down with such rectangular windowing of the DMT signal, as in Fig. 18. Asynchronous crosstalk may be such that especially with misaligned symbol boundaries that a really worstcase crosstalker could have its peak sidelobe aligned with the null of another tone (this is actually rare, but clearly represents a worst case). To confine this worst case, the methods of the next two subsections are used.

3.2.4.1. Windowing of Extra Suffix and Prefix. This section explains how windowing can be implemented without the consequence of intermodulation distortion when digital duplexing is used. Windowing in digitally duplexed DMT exploits the extra samples in the cyclic suffix and cyclic prefix beyond the minimum necessary. Since the cyclic extension is always fixed at L = 640samples in VDSL, there are always many extra samples. Figure 19 shows the basic idea — the extra suffix samples are windowed as shown with the extra extension samples now being split between a suffix for the current block and a prefix for the next block. The two are smoothly connected by windowing, a simple operation of timedomain multiplication of each real sample by a real amplitude that is the window height. The smooth interconnection of the blocks allows for more rapid decay in the frequency domain of the PSD, which is good for crosstalk and other emissions purposes. A rectangular window will have the per tone (baseband) rolloff function given by



pre	DMT symbol	suf		pr	е
			Extra suffix	Extra prefix	

Total cyclic extension is 640 samples

Figure 19. Illustration of windowing in extra suffix/prefix samples—smooth connection of blocks without affecting necessary properties for digital duplexing.

or the so-called sinc function in frequency. Clearly, a smoother window could produce a more rapid decay with frequency. A logical and good choice is the so-called raised-cosine window. Let us suppose that the extra cyclic suffix contains 2L' + 1 samples in duration, an odd number.⁹ Then, the raised cosine function has the following time-domain window (letting the sampling period be T'), where time zero is the first point in the extra cyclic suffix and the last sample being time 2L'T' and the centerpoint thus L'T':

$$W_{T,rcr}(t) = rac{1}{2} \left\{ 1 + \cos\left(\pi \cdot \left[rac{t}{L'T'}
ight]
ight)
ight\}$$
 $t = 0, \dots, L'T', \dots, 2L'T'$

One notes that the window achieves values 1 at the boundaries and is zero on the middle sample and follows a sampled sinusoidal curve in between. The points before time L'T' are part of the prefix of the current symbol, while the points after L'T' are part of the suffix of the last symbol.

The overall window (which is fixed at 1 in between) has Fourier transform (let $\alpha = L'/L - L'$), ignoring an insignificant phase term

$$W_T(f) = rac{L'}{lpha} \cdot rac{\sin\left(rac{\pi f}{f_0}
ight)}{\left(rac{\pi f}{f_0}
ight)} \cdot rac{\cos\left(rac{lpha \pi f}{f_0}
ight)}{1 - \left(rac{2lpha \pi f}{f_0}
ight)^2}$$

Larger α means faster rolloff with frequency. This function is improved with respect to the sinc function, especially a few tones away from an up/down boundary. Reasonable values of α corresponding to 100–200 samples will lead to even the peaks of the NEXT sidelobes below –140 dBm/Hz at about 200 kHz spacing below 5 MHz. The reduction becomes particularly pronounced just a few tones away, and so at maximum, a very small loss may occur with asynchronous crosstalk. Thus, signals other than 4.3125kHz DMT see more crosstalk, but within 200 kHz of a frequency edge, such NEXT is negligible. This observation is most important for studies of interference into home LAN signals like G.pnt, which at present almost certainly will not use 4.3125-kHz-spaced DMT.

3.2.4.1.1. Overlapped Transmitter Windows. Figure 20 shows overlapped windows in the suffix region. The smoothing function is still evident and some symmetrical windows (i.e., square-root cosine) have constant average power over the window and the effective length of



Figure 20. Illustration of overlapped windowing.

 9 If even, just pretend it is one less and allow for two samples to be valid duplexing endpoints.

the window above L' can be doubled, leading to better sidelobe reduction. This overlapping requires an additional 2L' + 1 additions per symbol, a negligible increase in complexity.

3.2.4.1.2. Receiver Windowing. Receiver windowing can also be used to again filter the extra suffix and extra prefix region in the receiver, resulting in further reduction in sidelobes. Figure 21 shows the effect on VDSL NEXT for both the cases of a transmitter window and both a transmitter and receiver window. Note the combined windows has very low transmit spectrum (well below FEXT in a few tones) and below -140 dBm/Hz AWGN floor by 40 tones. If it is desirable to further reduce VDSL NEXT to zero, an additional small complexity can be introduced as in the next subsection with the adaptive NEXT canceler.

3.2.4.1.3. Adaptive NEXT Canceler for Digital Duplexing. Figure 22 illustrates an adaptive NEXT canceler and its operation near the boundary of up and down frequencies in a digitally duplexed system. Figure 22 is the downstream receiver, but a dual configuration exists for the upstream receiver. Note that any small residual upstream VDSL NEXT left after windowing in the downstream tone n (or in tones less than n in frequency index) must be a function of the upstream signal extracted at frequencies n + 1, n + 2, ... at the FFT output. This function is a function of the frequency offsets between all the NEXTs and the VDSL signal. This timing clock offset is usually fixed, but can drift with time slowly. An adaptive filter can eliminate the NEXT as per standard noise cancellation methods [1]. A very small number of tones are required for the canceler per up/down edge if transmit windowing and receiver windowing are used. Adaptive noise cancellation can be used to make VDSL self-NEXT negligible with respect to the -140-dBm/Hz noise level. This allows full benefit of any FEXT reduction methods that may also be also in effect (note that the NEXT is already below the FEXT even without the NEXT canceler, but reducing it below the noise floor anticipates a VDSL system's potential ability to eliminate or dramatically reduce FEXT).

3.3. DMT VDSL Framing

The DMT transmission format supports Reed-Solomon forward error correction [1] and convolutional/triangular interleaving. The Reed-Solomon code is the same as that used for ADSL with up to 16 bytes of overhead allowed per codeword. There is no fixed relationship between symbol boundaries and codeword boundaries, unlike ADSL.

Instead any payload data rate that is an integer multiple of 64 kbps (implying an even integer multiple of payload bytes on average per symbol) can be implemented with dummy byte insertion where necessary and as described by Schelestrate [22]. Triangular interleaving that allows interleaving at a block length that is any integer submultiple of a codeword length (in bytes) is allowed (ADSL forced the block size of the interleaver to be equal to the codeword length). Given the high speeds of VDSL, the loss caused by dummy insertion is small, compared to the implementation advantage of decoupling symbol length from codeword length. Triangular interleaving was described by Starr et al. [1] and described again by Schelestrate [22].



Figure 21. PS = transmitter windowing (pulseshaping) and "window" here means receiver window. This simulation is for a 1000-m loop of 0.5-mm transmission line (24-gauge).



Latency can take on any value between 1 ms (fast buffer requirement) and 10 ms (slow buffer default) or more. The latency is determined according to codeword, data rate, and interleave depth parameter choices as in [22]. Fast and slow data is combined according to a frame format that no longer includes the synchronization symbol of ADSL, and has updates of the fast and slow control bytes with respect to ADSL. Superframes are no longer restricted to just 69 symbols as in ADSL.

3.4. Initialization

The various aspects of training of a DMT modem [1]. The VDSL training procedure is described in Ref. 22 and compatible with the popular "g.handshake" (g.994) methods of the ITU. The fundamental steps of training are the same as in [1] with the LT now being expected to set a timing advance and measure round trip delay of signals so that the digital-duplexing becomes automatic. The length of cyclic prefix versus suffix and other detailed framing parameters are set through various initialization exchanges.

One feature of digital duplexing is that it does allow very simple echo cancellation if there is band overlap. With synchronized symbols, there is only one tap per tone to do full echo cancellation where that may be appropriate. However, the NEXT generated by overlapping bands at high frequencies might discourage one from trying unless NEXT cancellation (coordinated transmitters and receivers) can also be used, which would also only be one tap per tone per significant crosstalker. The reader is referred to Ref. 22 for more details. **Figure 22.** Adaptive noise canceler for elimination of VDSL self-NEXT in asynchronous VDSL operation. Shown for one upstream/downstream boundary tone (which can be replicated for each up/down transition tone that has NEXT distortion with asynchronous VDSL NEXT). No canceler is necessary if NEXT is synchronous.

4. MULTIPLE-QAM APPROACHES AND STANDARDS

The VDSL system specified by Oksman [21] uses either CAP or QAM as a modulation scheme [1] and frequencydivision duplexing (FDD) to separate the upstream and downstream channels. There are two carriers or equivalently center frequencies, both with 20% excess bandwidth raised cosine transmission in each direction, following frequency plans 997 (Europe) or 998 (North America). The symbol rate of each of the signals is any integer multiple of 67.5 kHz, and the carrier/center frequencies can also be programmed as any integer multiple of 33.75 kHz. This allows for the receiver for each signal to estimate signal quality and request and appropriate center frequency and symbol rate, as well as corresponding signal constellation, which can be any integer QAM constellation from 4 points to 256 points as described in detail by Oksman [21]. Radio-frequency emission control occurs through programmable notch filters in the transmitter, for which a decision feedback equalizer in the receiver can partially compensate.

With overhead included, data rates are certain integer multiples (not all) of 64 kbps up to 51.84 Mbps downstream and 25.92 Mbps upstream.

4.1. Profiling in SCM VDSL

To accommodate short (<1 or 1000 ft), medium (1000– 3000 ft), and long (>3000 ft) transmission at both asymmetric and symmetric rates, SCM VDSL can transmit up to 4 QAM signals, two in each direction. For long loops, only one carrier downstream and one carrier upstream (just above the downstream band) is permitted. For medium range, a second downstream carrier is permitted in addition to the two carriers of long loops, while the short loops can use all four carriers. It was basically this 4-max carrier feature of SCM that forced the number of bands in the 997 and 998 frequency plans.

Figure 23 depicts the concepts of the profiles that are created by altering the center frequencies and symbol rates chosen. Notches at radio bands must be inserted to ensure emissions meet radio requirements, however there are not a sufficient number of carriers to simply achieve this notching by reversing direction. The 998 spectrum plan does leave one radio band as a reversal point near 4 MHz between upstream and downstream transmission. As Fig. 23 illustrates, larger symbol rates will likely be accompanied by larger center/carrier frequencies.

Decision-feedback equalization is presumed in the receiver and no Tomlinson precoding is used, even though FEC is used. The FECs in Ref. 21 and interleaving are basically identical to those used in the DMT standard. The DFE is described in the next subsection.

4.2. Operation of the DFE

The way a DFE handles RF interference (RFI) and notching is briefly discussed with reference to Fig. 24.

The analog front end (AFE) of any VDSL system needs to do some analog processing to reduce very strong RFI to an acceptable level and avoid overloading of the receiver's A/D and other input circuitry. This issue is not discussed any further here. The feedforward filter of the DFE creates a notch around the frequency at which the RF interferer

Asymmetric long profiles 998

is located, so that very little RFI is present at the output of this adaptive filter. The energy that is removed from the received signal by this notch is then restored by the feedback filter in such a way that the folded spectrum at the input of the slicer is flat. Actually, this is often claimed to be optimum performance by SCM advocates, but is not-optimum performance only can occur when the transmit band is silenced and more carriers are used to have a QAM signal on each side of the notch [25]. DFEs with multiple or single deep notches can require high precision implementation and must execute at the symbol rate, leading to billions of operations per second being required at VDSL speeds. Most QAM designers reduce the number of taps from the levels needed for excellent performance because of this complexity problem. Instead, designers hope that the difficult notching is not required often and then the small number of taps in the equalizer is sufficient.

In a way, SCM VDSL designers reduced system complexity in early chips by ignoring difficult channels and hoping that the low-complexity chips would be consequently attractive. QAM is suited well with the DFE on channels that have continuous transmission bands and mediocre distortion, where it can eliminate generation of multiple carriers. However, as the channel distortion grows, the complexity of the DFE quickly overtakes the complexity of the multiple carrier generation and QAM cannot handle channels with the severest distortion of VDSL well. SCM designers hope that an early low-cost solution can be replaced by increasing complex QAM components of the future that gradually address an increasing number of severe distortion situations in VDSL.



Figure 23. Basic concept of profiling in SCM VDSL.



Figure 24. Principle of operation of the DFE in the presence of RFI.

5. ETHERNET IN THE FIRST MILE (EFM)

Figure 26 shows the basic concept of EFM [38]. One, two, or four phone lines may be coordinated to deliver 10 or 100 Mbps symmetric VDSL service in EFM. The 10 mbps version is also recently known as MDSL. While "Ethernet" physical-layer copper twisted-pair standards have long been established, they are restricted to a length of 100 m for 10-Mbps (10base-T), 100-Mbps (100base-T) and 1000-Mbps (1000base-T). Each uses a category 5 set of 4 twisted pairs in synchronous point-to-point transmission. Longer-length transmission fundamentally requires a different physical-layer modulation, while the upper layer "Ethernet/TCP/IP" functionality can be maintained so that the DSL line essentially looks like a "long-range ethernet." Since the current physical-layer Ethernet often uses 3 or 4 twisted pairs in coordination, admitting that same possibility for the EFM versions (clearly longer length for any given speed can be attained by sharing the transmission bandwidth of several lines to achieve the desired rate of 10 or 100 Mbps) as well as the possibility of carrying the entire data rate on just one line also (over a distance shorter than 2 or 4 coordinated lines).

Presently, the IEEE 802.3 standards committee is studying EFM possibilities, and has selected VDSL for the transmission format. EFM appears interested in only symmetric transmission where VDSL under plans 997 and 998 are clearly designed for asymmetric transmission. However, the flexible VDSL spectra under the DMT standard in Section 3 clearly does allow different band use for EFM where appropriate.

Some documents on EFM line modeling have appeared [38,39], but models are not fully accepted nor standardized presently for long-length lines. In particular, FEXT modeling at high frequencies becomes very important, especially with the use of multiple lines at wide coordinated bandwidths.

5.1. Multiline FEXT Modeling

The multiple-input/multiple-output (MIMO) characterization of a cable of twisted pairs merits attention and measurement for studies in Ethernet in First Mile (EFM) efforts. As noted, groups of twisted pair within a cable may be combined for better transmission/duplexing: The interaction between lines within a subgroup or the entire cable can be exploited to improve performance and reduce transceiver complexity, motivating a model. Reference 39 suggests a temporary model for MIMO FEXT that can be used to evaluate/test EFM.

5.1.1. The MIMO FEXT Channel. Figure 25 illustrates the matrix or MIMO FEXT twisted-pair channel. Each of the M inputs to this matrix channel may produce a component of the signal at each of the K outputs. Usually, M = K. For instance, a quad (4 twisted pairs tightly packed together) has M = 4 inputs, K = 4 outputs, and a total of $16 = M \cdot K$ transfer functions of interest. These $M \cdot K$ transfer functions can be summarized in a $K \times M$ matrix



Figure 25. Matrix channel.



Figure 26. Ethernet in first mile illustration (shaded area is EFM interest) (courtesy of H. Barass).

H. The $K \times 1$ vector of channel outputs **Y** is then related to the $M \times 1$ vector of channel inputs **X** by **Y** = **HX**.

Ultimately, the designer would desire the exact **H** for each binder of wires: Any FEXT information is contained within this matrix. Approximate models are of interest in evaluating the various EFM opportunities in terms of range, rates, and service applications/market. In recognition that such transfer matrices are not well known, Reference 39 suggests an **H** model for temporary use in EFM studies. NEXT matrix models are of less MIMO interest since NEXT is either avoided by duplexing choice or by echo/NEXT cancellation between lines.

The *km*th element of $\mathbf{H} = [H_{km}(f)]_{k=1,\dots,K}$ is the transfer function from input *m* to output *k*. When m = k, H(f) is simply the transfer function of the *k*th line, $H_{kk}(f)$, and can be determined from basic transmission line theory, given the length and RLCG parameters of the line [10]. Reference 10 also models FEXT power transfer of the offdiagonal terms as proportional to the line transfer function $|H_{kk}(f)|^2$, the square of frequency f^2 , and the length of the line (in meters), *d*, which is explained on p. 90 of Ref. 11. This corresponds to a crosstalk–insertion loss transfer path of

$$H_{km}(f) = h_{\text{fext}} H_{kk}(f) \cdot (jf) \cdot \sqrt{d}$$
(1)

with a worst-case value of $h_{\text{fext}} = \sqrt{7.74 \times 10^{-21} \cdot (.3048 \text{ m/ft})} = 4.8 \times 10^{-11}$ for two adjacent category 3 phone company (telco) plant crosstalking lines. Equation (1) is for one crosstalking line — thus an extra multiplicative factor of $(K-1)_r^{0.6}$ used in models that average several lines is not used because that factor is for more than just two adjacent crosstalking lines. The factor h_{fext} is reduced nominally by a factor of 10 (or 20 dB) for category 5 wiring with tighter twisting. However, quads in the telephone plant

that instead sometimes twist all 4 lines in an ensemble, may have a value higher than the one above, as much as an increase by a factor of 20 dB. Thus, a range of h_{fext} may be given by

 $\begin{array}{l} ({\rm Category \ 5 \ independent \ twists}) \ 4.8 \times 10^{-12} \leq h_{\rm fext} \leq 4.8 \\ \times \ 10^{-10} \ ({\rm category \ 3 \ ensemble-twisted \ quads}). \end{array}$

This same FEXT model is often seen in a form that

describes only energy transfer, in other words, the squared magnitude of Eq. (1). Here, the model is converted to voltage transfer because EFM studies may desire phase information also. The equation above may sometimes be augmented by a linear phase term $e^{i2\pi f\tau}$, where τ is chosen to make the corresponding crosstalk impulse response causal. The matrix \mathbf{H} can then be formed by finding the insertion loss function for any twisted pair in the bundle and inserting this insertion loss along the diagonal terms of the matrix **H**. The off-diagonal terms are equal to Eq. (1) with possibly randomly chosen phase offsets and/or linear phase. Some schemes that coordinate the lines may find the details of the off-diagonal terms increasing important.¹⁰ In all cases, simple squaring of Eq. (1) and treating the FEXT like Gaussian noise (which is what current transceiver designs and implementations do) leads to the lowest possible (i.e., uncoordinated) performance.

However, actual individual FEXT insertion losses do not follow such a smooth characteristic with frequency

¹⁰ When the off-diagonal terms are significantly smaller than the diagonal, the best coordinated schemes all converge to make the line appear as if there were no FEXT. While the details of the transfer function are then not of consequence to performance, the implementation still depends on knowing the phase.



Figure 27. Illustration of 500-m example FEXT insertion loss functions (courtesy of John Cook of BTexact).

and indeed vary up and down with respect to the model in Eq. (1) (see Fig. 27). The variations can vary with the individual pairs as in Fig. 27.

The inaccuracy of the Eq. (*) model is increasingly evident at higher frequencies. There is a raised sinusoidal appearance to the magnitude transfer. This plot shows coupling functions between a pair on a 500 meter .5 mm cable with 50 pairs. Cosine terms can be added to the model to closely approximate the location of the dips and peaks in frequency seen in the measured FEXT insertion loss functions. Thus, the proposed model is

$$H_{km}(f) = \begin{cases} H(f) & m = k \\ n_{\text{lines}} \cdot \sqrt{h_{\text{fext}}d} \cdot (j2\pi f) \\ \cdot \left[1 + 0.3 \cdot \cos\left(\frac{2\pi fd}{c_{\text{line}}}\right) \\ -0.3 \cdot \cos\left(\frac{4\pi fd}{c_{\text{line}}}\right) \right] \cdot H(f) \cdot e^{j\varphi} & m \neq k \end{cases}$$

where H(f) is as derived above from standard transmission theory. c_{line} is the speed of light on the media (often just use 300 Mm/s, and φ is a phase term [i.e., $\varphi = 2\pi f \tau + \phi_{km}$, where τ makes the response causal and ϕ_{km} is chosen from a uniform distribution over $(0, 2\pi)$ independently for each pair of indices m and k]. Normalizing each offdiagonal entry by the factor $n_{\text{lines}} = (K-1)^{.6/2}/\sqrt{(K-1)} =$ $(K-1)^{-0.2}$ on average can account for the fact that distant lines have less crosstalk than close lines within the bundle and produces a slightly more accurate model when K = 25or 50.

The following table suggests values for h_{fext} and n_{lines} :

	Category 5 Quad	Category 3	Telco Quads
$h_{ m fext} \ n_{ m lines}$	$4.8 imes 10^{-12}\ 3^{-0.2}=0.803$	$\begin{array}{l} 4.8\times 10^{-11} \\ 49^{-0.2}=0.459 \end{array}$	$4.8 imes 10^{-10}\ 3^{-0.2}=0.803$

5.1.2. Summary. This proposed MIMO FEXT model attempts to augment well-known existing models to include

Sets of 4 lines—which may have better or worse coupling depending on the type of quad and associated independent (Cat 5) or ensemble (Cat 3 telco quad) twisting

Phase—phase coupling between lines that can be important for implementation of coordinated transmission schemes

Notches—the length-dependent frequency variation not included in earlier models, but that may become important in EFM studies

This model is in somewhat of a state of improvement at the time of writing and readers may want to pursue dynamic spectrum management and future EFM standards for any updates occurring after the time of writing. Reference 41 enumerates EFM data-rate vs range possibilities.

BIOGRAPHY

John M. Cioffi — BSEE, 1978, Illinois; PhDEE, 1984, Stanford; Bell Laboratories, 1978–1984; IBM Research, 1984–1986; EE Prof., Stanford, 1986–present. Cioffi founded Amati Com. Corp in 1991 (purchased by TI in

2806 VERY HIGH-SPEED DIGITAL SUBSCRIBER LINES (VDSLs)

1997) and was officer/director from 1991-1997. Currently he is on the boards or advisory boards of BigBand Networks, Coppercom, GoDigital, Ikanos, Ionospan, IteX, Marvell, Kestrel, Teknovus, Charter Ventures, and Portview Ventures, and a member of the U.S. National Research Council's CSTB. Cioffi's specific interests are in the area of high-performance digital transmission. Various Awards: Member, National Academy of Engineering 2001; IEEE Kobayashi Medal (2001), IEEE Millennium Medal (2000), IEEE fellow (1996), IEE JJ Tomson Medal (2000), 1999 University of Illinois Outstanding Alumnus, 1991 IEEE Comm. Mag. best paper; 1995 ANSI T1 Outstanding Achievement Award; NSF Presidential Investigator (1987-1992). Cioffi has published over 200 papers and holds over 40 patents, most of which are widely licensed, including basic patents on DMT, VDSL, and vectored transmission.

BIBLIOGRAPHY

- 1. T. Starr, J. M. Cioffi, and P. Silverman, *Understanding DSL Technology*, Prentice-Hall, Upper Saddle River, NJ, 1999.
- 2. W. Chen, DSL: Simulation Techniques and Standards Development for Digital Subscriber Lines, Macmillan Technical Publishing, March 1998.
- J. A. C. Bingham, ADSL, VDSL, and Multicarrier Modulation, Wiley, New York, 2001.
- D. A. Rauschmayer, ADSL/VDSL Principles: A Practical and Precise Study of Asymmetric Digital Subscriber Lines and Very High Speed Digital Subscriber Lines, Macmillan, 1998.
- 5. C. K. Summers, *ADSL: Standards, Implementation, and Architecture*, CRC Press, Boca Raton, FL, June 1999.
- P. S. Chow, J. S. Tu, and J. M. Cioffi, Performance evaluation of a multichannel transceiver for ADSL and VADSL services, *IEEE J. Select. Areas Commun.* 9(6): 909–919 (Aug. 1991).
- J. Cioffi and J. A. C. Bingham, A proposal for consideration of a VADSL standards project, ANSI Contribution T1E1.4/94-183, Dec. 5, 1994; see also J. Cioffi and K. Jacobsen, 15 Mbps on the Mid-CSA, ANSI Contribution T1E1.4/94-088, Palo Alto, CA, April 18, 1994, and Range/Rate Projections for NxDMT, ANSI Contribution T1E1.4/94-125, June 6, 1994.
- K. Foster, A proposal to study the feasibility of high-speed copper drop standardization, *British Telecom Contributions TD 56 and WD 62 to ETSI RG12 subgroup of TM3*, Rome, Italy, Oct. 10, 1994.
- 9. ATM Forum, ATM User-Network Interface Specification, V.3.0, Physical Layer UNI Interfaces STS-1 for 51.84 Mbps 100 meter transmission on unshielded twisted pair; see http://www.atmforum.com/atmforum/library/53bytes/ backissues/others/53bytes-0795-5.html, May 1995.
- DAVIC (Digital Audio Visual Council), Specification 1.4—Part 8 Lower-Layer Protocols and Physical Interfaces, http://www.ccett.fr/dam/dvc spec.htm, Short-range baseband asymmetrical PHY on copper and coax, ca., 1995.
- T. Starr, M. Sorbara, J. M. Cioffi, and P. Silverman, DSL Advances, Prentice-Hall, Upper Saddle River, NJ, 2002.
- 12. J. Singh, (CEO, OnFiber), Presentation on future of fiber transmission and broadband access, to Computer Science and Telecommunications Board of United States National Research Council, Palo Alto, CA, Jan. 26, 2001.

- IEEE Standard 802.11a-1999, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification: Highspeed Physical Layer in the 5 GHz Band, issued by IEEE Computer Society, Sept. 16, 1999.
- 14. International Telecommunications Draft Standard G.pnt, Study Group SG15/Q4, 2001.
- 15. American National Standard T1.417, Spectrum Management, 2001.
- T. Riley, Proposal to add a new work program to the VDSL project, T1E1.4 Contribution 99-, Clearwater, FL, Dec. 5, 1999.
- J. Cioffi, G. Ginis, and W. Yu, Performance of splitterless DSL at higher speeds, ANSI Contribution T1E1.4/99-558, Clearwater, FL, Dec. 5, 1999.
- ETSI TS 101 270-2 v1.1.5, (2000-12) TM, VDSL Part 2 Transceiver Specification, European Telecommunications Standards Institute Report on VDSL, TM (transmission and multiplexing), Access Transmission Systems on Metallic Access Cables, ETSI, Sophia Antipolis, 2000.
- J. Cioffi and W. Yu, G.vdsl: Robust VDSL in the presence of bridged taps, *ITU Contribution NG-039*, Nashville, TN, Nov. 2, 1999.
- Q. Wang, ed., Very-high-speed digital subscriber line (VDSL) metallic interface, Part 1: Functional requirements and common specification, ANSI Temporary Standard, T1.424-2002.
- V. Oksman, ed., Very-high-speed digital subscriber line (VDSL) metallic interface, Part 2: Single-carrier modulation specification, ANSI Temporary Standard, T1.424-2002.
- S. Schelestrate, ed., Very-high-speed digital subscriber line (VDSL) metallic interface, Part 3: Multicarrier modulation (MCM) specification, ANSI Temporary Standard, T1.424-2002.
- 23. J. Cioffi et al., North American Prioritization of high-speed asymmetric VDSL service with respect to other VDSL services in response to G.vdsl issue item 2.7, ANSI Contribution T1E1.4/99-393, Baltimore, MD, Aug. 23, 1999.
- FCC Order 00-336, Second Memorandum Opinion and Order, adopted Sept. 7, 2000; released Sept. 8, 2000, Ameritech and SBC Corporations.
- 25. V. Erceg et al., Channels or fixed wireless applications, *IEEE 802.16 Standards Contribution*, Xc-00/NNr0, Feb. 23, 2001, Tampa, FL; see also The IEEE 802.16 Working Group on Broadband Wireless Access Standards, fixed-wireless access standard and activities, Webpage *http://wirelessman.org/*, Jan. 16, 2001.
- ETSI EN 300 429 V1.1.2 (1997-08), Digital Video Broadcasting (DVB), Framing structure, Channel Coding, and Modulation for Cable Systems.
- M. Isaksson et al., Zipper—a flexible duplex method for VDSL, Proc. Int. Conf. Copper Wire Access Systems (CWAS '97), Budapest, Hungary, Oct. 1997, pp. 95–99; see also M. Isaksson et al., Zipper—a duplex scheme for VDSL based on DMT, ANSI T1E1.4/97-016, Feb. 3–7, 1997.
- K. Cheong et al., Soft cancellation via iterative decoding to mitigate the effect of home-LANs on VDSL, ANSI Contribution T1E1.4/99- 333R2, Baltimore, MD, Aug. 23, 1999.
- 29. J. Cioffi, G. Ginis, W. Yu, and C. Zeng, Example improvements of dynamic spectrum management, ANSI Contribution T1E1.4/2001-089R1, Costa Mesa, CA, Feb. 23, 2001.

- J. Cioffi et al., Construction of modulated signals from filter bank elements and equivalence of line codes, ANSI Contribution T1E1.4/99-395, Baltimore, MD, Aug. 23, 1999.
- G. Cherubini, E. Eleftheriou, S. Oelcer, and J. Cioffi, Filtering elements to meet requirements on power spectral density, *ANSI Contribution T1E1.4/99-429*, Baltimore, MD, Aug. 23, 1999.
- 32. J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney, Jr., MMSE decision feedback equalizers and coding: Parts I and II, *IEEE Trans. Commun.* 43(10): 2582–2604 (Oct. 1995).
- 33. J. M. Cioffi and G. D. Forney, Jr., Generalized decision-feedback equalization for packet transmission with ISI and Gaussian noise, in A. Paulraj, V. Roychowdhury, and C. Schaper, eds., *Communication, Computation, Control, and Signal Processing*, Kluwer, Boston, 1997 (a tribute to Thomas Kailath).
- F. Sjöberg et al., Zipper a duplex method for VDSL based on DMT, *IEEE Trans. Commun.* 47(8): 1245–1252 (Aug. 1999).
- D. J. G. Mestdagh, M. Isaksson, and P. Ödling, Zipper VDSL: A solution for robust duplex communication over telephone lines, *IEEE Commun. Mag.* 38(5): 90–96 (May 2000).
- F. Sjöberg et al., Asynchronous zipper, Proc. IEEE Int. Confer. Communications (ICC'99), Vancouver, Canada, June 1999, Vol. 1, pp. 231–235.
- F. Sjöberg et al., Performance evaluation of the zipper duplex method, *Proc. IEEE Int. Confe. Communications* (ICC'98), Atlanta, GA, June 1998, pp. 1035–1039.
- 38. IEEE 802.3 Draft, Ethernet in the First Mile Copper Standard Test and Data Rate Project, *IEEE 802.3 Standards Contribution*, Los Angeles, Oct. 18, 2001.
- J. Cioffi and J. Fang, A Temporary model for EFM/MIMO cable characterization, *IEEE 802.3 Standards Contribution*, Los Angeles, Oct. 18, 2001.
- ANSI Draft Dynamic Spectrum Management Report. ANSI T1E1.4/2002-R2, August, 2002, Westminster, Co.
- J. Cioffi, G. Ginis, and K. B. Song, "Coordinated Level 2 DSM Results: Vectoring of multiple DSLS," ANSI Contribution TlE1.4/2002-059, Vancouver, BC, Feb. 18 2002.
- 42. J. Cioffi, J. Lee, and S. T. Chung, "10 MDSL Beyond all Goals, and Spectrally Compatible with ADSL and VDSL, From Co or RT," ANSI Contribution TlE1.4/2002-129, Atlanta, GA, April 8, 2002.

VIRTUAL PRIVATE NETWORKS

MARC-ALAIN STEINEMANN TORSTEN BRAUN MARC DANZEISEN MANUEL GÜNTER University of Bern Bern, Switzerland

1. INTRODUCTION

A *virtual private network* (VPN) is a private network constructed by public lines or connections using secure methods to transfer information. For example, VPN technology allows organizations to securely extend their network services across shared public networks such as the Internet to remote users, branch offices, and partner companies.

Large corporations used to interconnect local headquarters and branch offices with leased connections provided by telecommunication companies and ran private networks, so-called corporate networks. With the rise of the Internet technology more and more corporate networks switched from various networking protocols such as Novell to the TCP/IP protocol suite. Such private networks based on Internet technology are also referred to as *intranets*. Since leased lines are expensive and the corporations often already have Internet connectivity, there is an economic incentive to replace the expensive leased connections and to use the wide-area interconnectivity of the global Internet instead. However, two basic problems must be emphasized:

- 1. The intranet may use private addresses that are not unique in the global Internet and thus not routable [1].
- 2. The Internet protocol version 4 does not assure transmission privacy. While IP packets travel through the public Internet they may be viewed or even altered by third parties.

2. DIFFERENT TYPES OF VPNs

2.1. Subnet-to-Subnet and Access VPNs

Virtual private networks [2-4] encapsulate the packets with private addresses into packets with public addresses. This process is called *tunneling*. If privacy and authenticity of the encapsulated packets are desired, these can be ensured with cryptographic means.

Figure 1 shows the two most prominent VPN types: subnet-to-subnet VPNs and access VPNs. The subnetto-subnet VPN interconnects geographically distributed



Figure 1. Virtual private network types.

private IP subnets. All traffic leaving one subnet destined for another one is tunneled through the public Internet. The access VPN allows roaming users to dial into the virtual network from their home computers or via an arbitrary Internet point of presence (POP).

Figure 1 also illustrates the tunneling mechanism. It shows the structure of a tunneled IP packet originating from an application that runs within the private subnet X. The packet's destination is a computer in a remotely located part of the VPN (the private subnet Y). The subnets X and Y use private IP addresses that cannot be routed in the public Internet. The address structure of the VPN is invisible from the outside. The access routers of subnets X and Y incorporate VPN functionality. They have an interior network interface with a private IP address and an exterior network interface with a public IP address. The access router at X recognizes that the packet in question must be tunneled. It knows the public interface of the access router of subnet Y and uses that address as the destination address and its own public address as the source address. The access router (also referred to as the *tunnel endpoint*) creates a new IP packet with these new addresses and puts the original packet into the payload of the new packet. The payload is then encrypted. The new packet is sent to the tunnel endpoint at Y. There, the router extracts the payload of the packet and decrypts the content. In this manner, the original packet is restored and can be routed on the private subnet Y toward the originally intended destination.

The access VPN case also uses tunnels. However, there are two distinct possibilities. Either the home PC acts as a tunnel endpoint or the POP of an Internet service provider (ISP) acts as tunnel endpoint.

While a VPN may be useful for a small-to-medium-sized company, the management of the VPN would require additional equipment and personnel. As a consequence, there exists a market for VPN services that lets the customers outsource the management of their VPN. The ISP can deploy VPN capable border routers and use them to introduce a VPN on-demand service [5]. Thereby, several VPNs can be managed on the same infrastructure by the same personnel (ISP staff) so that both the customer and the provider can profit from the economy of scale.

2.2. Encapsulation

Today, many different types of VPN technologies exist such as layer 2 VPNs based on frame relay and asynchronous transfer mode (ATM) networks; remote-access VPNs such as PPTP and L2TP; and IPSec-based VPNs.

2.2.1. Link-Layer VPNs (Layer 2). The Integrated Services Digital Network (ISDN), frame relay and asynchronous transfer mode are connection-oriented networks on link level (layer 2) that support the establishment of link-layer VPNs. Nowadays, most link-layer VPNs are established by frame relay and ATM technology. IP network links over these underlying connection-oriented network technologies are based on overlay models. In this case, meshes of connections have been established to interconnect IP routers of particular VPNs by providing a tunneling infrastructure.

Other but similar types of virtual networks based on link-level mechanisms are virtual local-area networks (VLANs) that can be established using IEEE 802.1Q, ATM LAN emulation (LANE), or multiprotocol over ATM (MPOA) technologies.

A major disadvantage of layer 2 VPNs and also VLANs is the need for a homogeneous topology throughout the entire VPN and the complexity involved in managing two different network technologies, namely, IP and the underlying network technology, for a single VPN. An advantage lies in the connection-oriented structure of those technologies. Links stay established and the tunneled packets follow the link and don't need to be routed as in IP-based VPNs. In addition, quality of service (QoS) is often provided implicitly by the connection-oriented network technologies.

2.2.2. Network-Layer VPNs (Layer 3). In contrast to the link-layer VPNs, where the location-independent IP provides layer 3 addresses and the location-dependent addresses are provided by layer 2 technology, in network layer VPNs, IP provides location-independent as well as the location-dependent addressing. For example, in a link-layer VPN, the location-independent IP addresses can be chosen by the user and the fixed medium-access channel (MAC) addresses are delivered by the network interface. In a network-layer VPN, the location-dependent IP addresses are provided by the intranet and the location-independent IP addresses are provided by the intranet and the location-independent IP addresses are provided by the VPN. VPNs based on tunneling mechanisms that use network-layer protocols such as IP or MPLS as outer header are called *network-layer VPNs*.

Tunneling (also called *packet encapsulation*) is a method of wrapping a packet into a new one by prepending a new header. The whole original packet becomes the payload of the new one. At the tunnel endpoints (usually border routers) the header is added (respectively removed) and the result is then forwarded again. Tunneling is often used to transparently transport packets of one network protocol through a network running another protocol.

IP VPN tunneling mechanisms often encapsulate IP packets into IP packets. This tunneling method is called *IP in IP* encapsulation (IPIP). With IPIP encapsulation encryption can be applied to the inner packet by using IPSec protocols.

Generic routing encapsulation (GRE) is another popular tunneling method. GRE is a multiprotocol carrier protocol. With GRE a router at each VPN site encapsulates protocolspecific packets in an IP header, creating a virtual pointto-point link to routers at other ends of an IP cloud, where the IP header is stripped off. By connecting multiprotocol subnetworks in a single-protocol backbone environment, IP tunneling allows network expansion across a singleprotocol backbone environment. GRE tunnels do not provide true confidentiality (no encryption functionality) but can carry encrypted traffic. It is possible to encapsulate almost every existing network protocol in GRE.

Standard protocols such as the Point to Point Tunneling Protocol (PPTP) and Layer 2 Forwarding (L2F) are required for supporting remote VPN access by single end systems. The protocols establish virtual point-to-point links between an end system and a VPN server. The VPN server acts as an interface of a VPN for remote end systems. The protocols mentioned above can carry any other network protocol and are themselves encapsulated in IP. PPTP and L2F have been developed further resulting in a standard called Layer 2 Tunneling Protocol (L2TP).

With multiprotocol label switching (MPLS) routing is independent from the destination address in the encapsulated packet. This independence from the routing decision and the destination address is obtained by establishing a label-switched path (LSP) instead of establishing an IP tunnel between the two routers of a common VPN. MPLS allows setting up tunnels by appending a MPLS header in front of the IP header. This 32-bit MPLS header avoids the large overhead by another IP header as it is required with IP-in-IP tunneling. Multiple MPLS headers are possible; thus labels can be stacked onto each other. Label stacking supports hierarchical tunnels and is in particular being used when building MPLS-based VPNs.

In a typical MPLS VPN scenario as shown in Fig. 2, a packet is classified at an ingress router of an ISP based on the incoming port number as belonging to a particular VPN. The ingress router has learned via Boarder Gateway Protocol (BGP) to which VPN it belongs, to which egress router the packet must be sent, and via which egress interface the destination is reachable. The ingress router appends two labels to a packet belonging to a VPN; the inner label specifies the egress port at the ISPs egress router, namely, the link toward the destination subnetwork of the VPN. The outer label is being used to forward the packet toward the egress router and can be learned by MPLS signaling protocols such as Constraintbased Routing (CR) using Label Distribution Protocol CR-LDP or Traffic Engineering Resource Reservation Setup Protocol (TE-RSVP). Both labels are popped by the egress router (edge router). Figure 2 shows an example VPN/MPLS scenario with a label-switched path (LSP) set up between ingress and egress of an ISP. This LSP is set up along the path and carries the traffic between the VPN subnets. Note that MPLS makes the private VPN addresses of a customer transparent to the routers of the ISP and that MPLS does not provide security mechanisms as IPSec does.

2.2.3. Firewalls and VPNs. VPN tunnels are initiated and terminated mainly by specially equipped routers equipped with the respective hardware and software for

establishing VPNs. If the organization at the endpoint of a tunnel needs additional security, the router can be replaced by a firewall router. It is also possible to establish VPNs through firewalls, that is, to tunnel a VPN link through a firewall. In the case of opening a firewall for a VPN tunnel, the instance allowing access to systems behind their firewall has to make sure that the other side deploys at least the same security policy level. If a host establishes a single unprotected connection to the Internet, and is at the same time connected through a VPN to computers behind a firewall, hackers can break in quite easily.

3. SECURITY AND THE INTERNET PROTOCOL

There exists a wide spectrum of technologies securing Internet communication, but most of them are dedicated to specific software applications. In that case, security is provided by the application layer. Good examples are Pretty Good Privacy (PGP) for mail encryption and browser-based authentication as well as Secure Sockets Layer (SSL) for traffic encryption between Web browser and Web server. These restrictions are not consistent with the requests of a large enterprise, and the average ISP that may never know precisely the kind of applications running tomorrow over today's networks.

3.1. Possible Threats on the Internet

VPNs are driven by security threats in the network environment and must fulfill three fundamental requirements:

- *Authentication*—the communicating persons must really be the persons they claim to be.
- *Confidentiality and privacy* no one shall be able to electronically eavesdrop traffic.
- *Integrity* the received traffic must not be altered in any way during transmission.

3.1.1. Spoofing. In IP networks it is difficult to know where information really originates. An attack called IP "spoofing" takes advantage of this weakness. Since the source IP address of a packet has no influence on routing, it can easily be forged. In this type of attack, a packet coming from one machine appears to be coming from another one. In fact, an IP source address is not trustable.



Figure 2. Different VPNs tunneled with MPLS over the same link.

3.1.2. Session Hijacking and "Man in the Middle" Attack. Spoofing makes it possible to take over a connection. Even initial authentication for each communication is no protection against session hijacking. A hacker can take over a session and stay invisible in the middle, pretending to be the respective peer of the two original session partners. The hacker thereby possibly filters and modifies all packets of the session. Identifying the communicating person once does not ensure that it remains the same person throughout the rest of the session. Each data source has to be authenticated throughout the whole session.

3.1.3. Electronic Eavesdropping. A large part of most networks is based on Ethernet LANs. This technology has the advantage of being cheap, universally available, and easy to expand. But it has the disadvantage of making sniffing easy. An even more severe situation nowadays exists in wireless LANs.

In Ethernet networks, every node can read each packet. Conventionally, each network interface card listens and responds only to packets specifically addressed to it. But it is easy to force these devices to collect every packet that passes on the wire. Physically, there is no way to detect from elsewhere on the network which network interface card is working in the so-called promiscuous mode.

Diagnostic tools called "sniffers" get the information out of the collected packets. Such tools can record all the network traffic and are normally used to quickly determine what is happening on any segment of the network. However, in the hands of someone who wants to listen in on sensitive communications, a sniffer is a powerful eavesdropping tool.

The grown Internet structure with the global backbones makes electronic eavesdropping on routers and especially on backbone routers very efficient. Also in virtual LANs that transfer clear text, packets can be eavesdropped easily.

3.2. The Security Architecture for the Internet Protocol (IPSec)

The Internet Engineering Task Force (IETF) standardized IP version 6 (IPv6) [6] to solve pending problems such as address shortage of the current version of the IP protocol (IPv4). A spinoff development of this process was the IP security architecture (IPSec), which introduces per packet security features. While the IP version 6 deployment has been delayed, the security architecture has been adopted by the current IP version (IPv4). A key motivation for this was that IPSec includes all security mechanisms needed to implement VPNs.

The Internet security architecture consists of a family of protocols. IPSec describes IP packet header extensions and packet trailers that provide security functions. The per packet security functions come from two protocols: the Authentication Header (AH) [7], which provides packet integrity, and authenticity and the Encapsulating Security Payload (ESP) [8], which provides privacy through encryption. AH and ESP (Figs. 3–5) are independent protocols that can be used separately and that can be combined. One reason for the separation was that there are countries that have restrictive regulations on



Figure 3. IPSec; IP packet after applying ESP in tunnel mode.



Figure 4. IPSec; IP packet after applying ESP in transport mode.

Original IP header	AH header	TCP header (or UDP or ICMP)	Data
-----------------------	--------------	--------------------------------------	------

Figure 5. IPSec; IP packet after applying AH in transport mode.

encrypted communication. There, IPSec can be deployed solely using AH because authentication mechanisms are not regulated.

The set of AH and ESP is required in order to guarantee interoperability between different IPSec implementations. Both protocols are specified independently of cryptographic algorithms. A new encryption algorithm, for example, can easily be added to IPSec. Both AH and ESP assume the presence of a secret key. This key material may be installed manually. A better and more scalable approach is to use the third protocol of the IPSec family: the Internet Key Exchange protocol (IKE) [9], described below.

3.2.1. The Encapsulation Security Payload. The Internet Assigned Numbers Authority (IANA) has assigned the protocol number 50 for the IPSec encapsulation security payload. ESP ensures privacy of the IP payload. For that purpose an ESP header and an ESP trailer clamp the IP payload between them. The payload and the trailer are encrypted. The ESP also provides optional authentication. Figure 6 depicts the ESP part of an IP packet transformed by ESP in transport mode. The ESP header is located after the IP header and contains the security parameter index to identify the security association. Furthermore, there is a sequence number that is incremented for each packet. This helps to detect replay attacks, where the attacker records a packet and resends it later.

The ESP trailer is added after the payload. The trailer includes padding that is necessary because the encryption



Figure 6. ESP part of an IPSec packet in transport mode.

algorithms often require the payload to be blocks of fixed length (e.g., 8 bytes). The pad length field encodes the length of the padding in bits. The next header field contains the protocol number of the next (eventually higher layer) protocol in the payload (e.g., IP or a concatenated IPSec protocol). Note that the trailer up to here is also encrypted. So, an attacker can, for example, not read what protocol is in the payload data. The ESP trailer may end with optional authentication data. The authentication data consist in a message authentication code (MAC) computed by a secure hash function. The input of the hash is a secret key, the ESP header, the ESP payload, and the rest of the ESP trailer. The MAC does not protect the initial IP header.

ESP supports nearly any kind of symmetric encryption. The default standard built into ESP, which assures basic interoperability, is 56-bit DES. ESP also supports some authentication (as does AH—the two options have been designed with some overlap).

3.2.2. The Authentication Header. The IANA has assigned the protocol number 51 for the IPSec authentication header. AH authenticates the packet so that a receiving IPSec peer can know for sure that the packet originates from the sending peer. Furthermore, the packet integrity is guaranteed. The receiver can verify that nobody has changed the packet while it was in transit between the peers. AH ensures this by calculating authentication data with a secure one-way hash function. The calculation also includes the secret key.

An attacker not knowing this key is neither able to forge a valid packet nor to authenticate the packet. Figure 7 depicts the AH part of an IP packet transformed by AH in transport mode. The AH header includes the next header field and encodes the payload length. The length is necessary because the authentication data are variable in length. The AH header, just like the ESP header, contains a security parameter index and a sequence number. Finally, there is the authentication data (the secure hash value). The authentication of AH also covers the original IP header, in contrast to the optional authentication of ESP. However, some fields of the IP header are excluded



Figure 7. AH part of an IPSec packet.

from the authentication, because their values may change during the forwarding of the packet. These exceptions are the time-to-live field that is decremented by each router and the Differentiated Services Code Point (DSCP) protocol.

The design of the authentication header protocol makes it independent from the higher-level protocol. It can be used with or without ESP. The different fields of the AH are

- The next header field that specifies the higher-level protocol following the AH.
- The pad length field is an 8-bit value specifying the size of the AH.
- The reserved field is reserved for future use and is currently always set to zero.
- The SPI identifies a set of security parameters to be used for this connection.
- The sequence number is incremented for each packet sent with a given security parameter index (SPI).
- Finally, the authentication data are the actual integrity check value (ICV), or digital signature, for the packet. It may include padding to align the header length to an integral multiple of 32 bits (in IPv4) or 64 bits (in IPv6).

To guarantee minimal interoperability, all IPSec implementations must support at least HMAC-MD5 (Keyed-Hash Message Authentication Code for the Message Digest 5 Algorithm) and HMAC-SHA-1 (Keyed-Hash Message Authentication Code for Secure Hash 1 Algorithm) for AH. IPSec, including AH and ESP, has been designed for both IPv4 and IPv6.

3.2.3. Transport and Tunnel Mode. Both ESP and AH have two modes: the transport mode and the tunnel mode. Transport mode just encrypts and authenticates the payload and a part of the IP header. It extends the IP headers by adding new fields. Transport mode allows the user to run IPSec from end-to-end (Fig. 8), while the tunnel mode is ideal for implementing a VPN tunnel at Internet access routers (Fig. 9).

The tunnel mode adds a complete new IP header (plus extension fields). In tunnel mode both AH and ESP can be



Figure 8. Transport mode.

Figure 9. Tunnel mode.

used to implement IP-VPN tunnels. AH and ESP dispose of a small standardized set of cryptographic algorithms to ensure authenticity and privacy. Tunneling takes the original IP packet and encapsulates it within the ESP. Then it adds a new IP header to the packet containing the address of the IPSec gateways. This mode allows passing nonroutable IP addresses or other protocols through a public network as the addresses of the inner header are hidden. Privacy is also given by hiding the original network topology.

3.2.4. Security Association and Security Policy Database. At some point in the network, both AH and ESP perform a transformation to IP packets. The IPSeccompliant nodes always form sender-receiver pairs, where the sender performs the transformation and the receiver reverses it. The relation between sender and receiver is described as a security association (SA). Note that the security association describes just one transformation and its inverse. Concatenated AH and ESP transformations are described by concatenated SAs. SAs can be seen as descriptions of "open" IPSec connections. Both IPSec peering machines store representations of security associations.

Under IPSec, the SA specifies the mode of the authentication algorithm used in the AH and the keys of that authentication algorithm. Also, it specifies the ESP encryption algorithm mode and the respective keys, the presence and size or absence of any cryptographic synchronization to be used in that encryption algorithm, how to authenticate traffic (protocols, encrypting algorithm and key), how to make communication private (again, algorithm and key), how often those keys need to be changed and the authentication algorithm, mode and transform for use in ESP, and the keys to be used by that algorithm. Finally it specifies the key lifetimes, the lifetime of the SA itself, the SA source address, and a sensitivity-level descriptor.

A SA is uniquely identified by a triple consisting of a security parameter index (SPI) (a 32-bit number), the destination IP address, and the IPSec protocol (AH or ESP). The sending party writes the SPI into the appropriate field of the IP protocol extension. The receiver uses this information to identify the correct security association. In that way the receiver is able to invert the transformation and to restore the original packet. Each IPSec-compliant machine may be involved in an arbitrary number of security associations.

Accordingly, a SA is a management construct used to enforce a security policy in the IPSec environment. The policy specifications are stored locally in every IPSec node's security policy database (SPD), which is consulted each time when processing inbound and outbound IP traffic, including non-IPSec traffic. The SPD contains different entries for inbound and outbound traffic. The SPD determines if traffic must be encrypted or can remain clear text or if traffic must be discarded. If traffic is encrypted, the SPD must point to the respective SA by a selector, a set of IP and upper-layer protocol field values to map traffic to a policy.

3.2.5. The Internet Key Exchange Protocol. If two parties would like to communicate using authentication and encryption services, they need to negotiate the protocols, encryption algorithms, and keys to use. Afterward they need to exchange keys (this might include changing them frequently) and keep track of all these agreements.

The Internet Key Exchange (IKE) protocol allows two nodes to securely set up a security association by allowing



Figure 10. IKE main mode: (a) first step; (b) second step; (c) third step.

these peers to negotiate the protocol (AH or ESP), the protocol mode, and the cryptographic algorithms to be used. Furthermore, IKE allows the peers to renew an established security association.

IKE uses the Internet Security Association and Key Management Protocol (ISAKMP) [10] to exchange messages. ISAKMP provides a framework for authentication and key exchange but does not define a particular key exchange scheme. IKE uses parts of the key exchange schemes Oakley [11] and SKEME [12].

IKE operates in two phases. In phase 1 the two peers establish a secure authenticated communication channel (also called *ISAKMP security association*). In phase 2 security associations can be established on behalf of other services (most prominently IPSec security associations). Phase 2 exchanges require an existing ISAKMP SA. Several phase 2 exchanges can be protected by one ISAKMP SA, and a phase 2 exchange can negotiate several SAs on behalf of other services.

ISAKMP SAs are bidirectional. The following attributes are used by IKE and are negotiated as part of the ISAKMP SA: encryption algorithm, hash algorithm, authentication method, and initial parameters for the Diffie-Hellman algorithm [13].

3.2.5.1. Phase 1 Exchange. IKE defines two modes for phase 1 exchanges: main mode and aggressive mode. The *main mode* consists of three request-response message

pairs. The first two messages negotiate the policy (e.g., authentication method) (Fig. 10a); the next two messages exchange Diffie-Hellman public values and ancillary data necessary for the key exchange (Fig. 10b). The last two messages authenticate the Diffie-Hellman exchange (Fig. 10c). The last two messages are encrypted and conceal the identity of the two peers.

The aggressive mode of phase 1 consists of only three messages (Fig. 11). The first message and its reply negotiate the policy. Moreover, they exchange Diffie-Hellman public values, ancillary data necessary for the key exchange, as well as identities. In addition the second message authenticates the responder. The third message authenticates the initiator and provides a proof of participation in the exchange. The final message may be encrypted. Aggressive mode securely exchanges authenticated key material and sets up an ISAKMP SA, but it reveals the identities of the ISAKMP SA peers to eavesdroppers. Note, that the choice of the authentication method influences the specific composition of the payload of this exchange. Note also, that IKE assumes security policies that describe what options can be offered during the IKE negotiation.

3.2.5.2. Phase 2 Exchange. A phase 2 exchange negotiates security associations for other services and is protected (encrypted and authenticated) based on an existing ISAKMP security association. The payloads of all phase 2



Figure 11. IKE aggressive mode.

messages are encrypted. A phase 2 exchange consists of three messages. The initiator sends a message containing a hash value, the proposed security association parameters and a nonce. The hash value is calculated over ISAKMP SA key material and proves authenticity.

The nonce prevents replay attacks. Optionally, the initial message can also contain key exchange material. Such optional phase 2 key exchange generates key material that is independent from the key material of the ISAKMP SA. If the new SA should be broken, the ISAKMP SA is thus not compromised. The initial message may also contain identifiers in case the new SA is to be established between peers different from the ISAKMP SA peers.

The responder replies with a message of the same structure as the initial message: an authenticating hash value, the selected SA parameters, and a nonce. If the initial message contained optional parameters, then these are also part of the reply. Finally, the initiator acknowledges the exchange with a third and final message containing yet another hash value.

3.2.5.3. Authentication. IKE establishes authenticated keying material. IKE supports four authentication methods to be used in phase 1: preshared secret keys, two forms of authentication with public key encryption, and digital signatures. Today's IKE implementations support X.509 certificates. Two computers not knowing each other can initialize a security association through the help of the commonly trusted third party that verified the certificates.

4. OUTLOOK

The move from legacy-technology-based VPNs such as frame relay and ATM to IP-based VPNs will go on and thereby accelerate the deployment of newer VPN techniques such as generalized MPLS (G-MPLS). GMPLS is being considered as an extension to the MPLS framework to include optical, non-packet-switched technologies. A more recent traffic engineering technology development in the context of G-MPLS is multiprotocol lambda switching (MP λ S). The major difference lies in the replacement of the traditional numeric MPLS labels by wavelengths (lambda).

Another trend are mobile devices. Mobile users, as described above, move around and connect through fixed wire dialup lines for example. These users are called "nomadic" users because the from the IP network view they remain locally immobile during a connection time. Roaming users that connect by mobile IP (MIP) require special solutions in the VPN area as with each handover of the mobile node the VPN tunnels need to be reestablished within very short timeframes. An interesting combination of the IPSec suite and the MIP protocols has been described [14] in which, where mobile hosts are allowed access to VPNs that are protected by firewalls from the public Internet.

BIOGRAPHIES

Manuel Günter received his Diploma Degree and his Ph.D. from the University of Berne (Switzerland) in 1998 and 2001, respectively. His research interests are new IP services, network and service monitoring, software engineering, and mobile agents. He now works as a scientific consultant for the Swiss National Bank.

Marc Danzeisen received his Diploma Degree from the University of Bern (Switzerland) in 2001. Since then he has been working as a Ph.D. student in the area of mobile ad hoc networks in collaboration with Swisscom Innovations AG (Switzerland). He is also active in the creation of future mobile network services.

Marc-Alain Steinemann received his Diploma Degree from the University of Bern (Switzerland) in 2000. He is now working as a Ph.D. student in the Computer Networks and Distributed Systems Research Group of the University of Bern. His research interests are remote learning and communication systems for the next-generation Internet.

Torsten Braun received his Diploma Degree and Ph.D. degree from the University of Karlsruhe (Germany) in 1990 and 1993, respectively. From 1994 to 1995 he has been a Guest Scientist with INRIA Sophia-Antipolis (France). From 1995 to 1997 he worked at the IBM European Networking Center Heidelberg (Germany), and at the end of that period served as a project leader and senior consultant. Since 1998, he has been a full Professor of Computer Science at the Institute of Computer Science and Applied Mathematics, heading the Computer Networks and Distributed Systems research group. He is a member of several international conference and workshop program committees and the foundation council of SWITCH (Swiss National Research Network).

BIBLIOGRAPHY

1. Y. Rekhter et al., Address Allocation for Private Internets, RFC 1918, Feb. 1996.

- 2. P. Ferguson and G. Huston, What is a VPN—part I, Internet Protocol J. 1(1): 2–11 (1998).
- 3. P. Ferguson and G. Huston, What is a VPN part II, *Internet Protocol J.* 1(2): 2–18 (1998).
- 4. B. Gleeson et al., A Framework for IP Based Virtual Private Networks, RFC 2764, Feb. 2000.
- I. Khalil, T. Braun, and M. Günter, Management of quality of service enabled VPNs, *IEEE Commun. Mag.* 39(5): 90–98 (May 2001).
- 6. S. Deering and R. Hinden, Internet Protocol, Version 6 (IPv6) Specification, RFC 2460, Dec. 1998.
- S. Kent and R. Atkinson, *IP Authentication Header*, RFC 2402, Nov. 1998.
- 8. S. Kent and R. Atkinson, *IP Encapsulating Security Payload* (*ESP*), RFC 2406, Nov. 1998.
- 9. D. Harkins and D. Carrel, *The Internet Key Exchange (IKE)*, RFC 2409, Nov. 1998.
- D. Maughhan, M. Schertler, M. Schneider, and J. Turner, Internet Security Association and Key Management Protocol, RFC 2408, Nov. 1998.
- H. Orman, The Oakley Key Determination Protocol, RFC 2412, Nov. 1998.
- F. Knabe, An overview of mobile agent programming, in Analysis and Verification of Multiple-Agent Languages, Vol. 1192 of Lecture Notes in Computer Science, Springer, June 1996; 5th LOMAPS Workshop.
- 13. B. Schneier, Applied Cryptography, Wiley, New York, 1996.
- 14. M. Danzeisen and T. Braun, Access of mobile IP users to firewall protected VPNs, Workshop on Wireless Local Networks at 26th Annual IEEE Conf. Local Computer Networks (LCN'2001), Tampa, FL, Nov. 15–16, 2001.

VITERBI ALGORITHM

A. J. VITERBI Viterbi Group San Diego, California

1. FUNDAMENTALS

The Viterbi algorithm is a computationally efficient technique for determining the most probable path taken through a Markov graph. The graph and underlying Markov sequence is characterized by a finite set of states $\{S_0, S_1, \ldots, S_n, \ldots\}$, state-transition probabilities $\Pr(S_j \to S_i)$, and the output (observable parameter) probabilities $p(y/S_j \to S_i)$ for all i and j, where the observables y are either discrete or continuous random variables. An example of a four-state Markov graph is shown in Fig. 1, where only the nonzero probability transitions are shown. Thus, for example, from S_1 the only nonzero transition probabilities are those to S_2 and S_3 , while from S_3 they are those to itself, S_3 , and to S_2 .

It is also convenient for the description of the algorithm to view the multistep evolution of the path through the graph by means of a multistage replication of the Markov graph known as a *trellis* diagram. Figure 2 is the trellis diagram corresponding to the four-state Markov graph of Fig. 1.



Figure 1. Markov graph example.



Figure 2. Trellis diagram for Markov graph of Fig. 1.

At the top of Fig. 2 are shown the successive observables $y(1), y(2), \ldots, y(k), \ldots$, each of which may be vectors corresponding to multiple observations per branch. Henceforth we use the notation y(k) to denote the observable(s) for the kth successive branch. Similarly, S(k) will denote any state at the kth successive node level (and we shall dispense with subscripts until necessary).

The goal, then, is to find the most probable path through the trellis diagram. A fundamental assumption is that successive Markov state probabilities $\Pr[S(k-1) \rightarrow S(k)]$ are mutually independent for all k, as are the conditional output probabilities $p[y(k)/S(k-1) \rightarrow S(k)]$. For any given path from the origin (k = 0) to an arbitrary node n, S(0), S(1), ..., S(n), the relative path probability (likelihood function) is given by

$$L = \prod_{k=1}^{n} \Pr[S(k-1) \rightarrow S(k)] p[y(k)/S(k-1) \rightarrow S(k)]$$

For computational purposes it is more convenient to consider its logarithm, which is given by the sum

$$\ln(L) = \sum_{k=1}^{\infty} m[y(k); S(k-1), S(k)]$$

where

$$m[y(k); S(k-1), S(k)] = \ln\{\Pr[S(k-1) \to S(k)] + \ln p[y(k)/S(k-1) \to S(k)]\}$$

which is denoted the *branch metric* between any two states at the (k - 1)th and kth node levels. We next define the *state metric*, $M_K(S_i)$, of the state $S_i(K)$ to be the maximum over all paths leading from the origin to the *i*th state at the Kth node level. Thus, again inserting subscripts where necessary, we obtain

$$\begin{split} M_{K}(S_{i}) &= \max_{\text{all paths } S(0), S(1), \dots, S(K-1)} \left\{ \sum_{k=1}^{K-1} m[y(k); S(k-1), S(k)] \right. \\ &+ m[y(K); S(K-1), S_{i}(K)] \right\} \end{split}$$

It then follows that to maximize this sum over K terms, it suffices to maximize the sum over the first K - 1 terms for each state $S_j(K - 1)$ at the (K - 1)th node and then maximize the sum of this and the Kth term over all states S(K - 1). Thus

$$M_K(S_i) = \max_{S_i(K-1)} \{M_{K-1}(S_j) + m[y(K); S_j(K-1), S_i(K)]\}$$

This recursion is known as the Viterbi algorithm. It is most easily described in connection with the trellis diagram. If we label each branch (allowable transition between states) by its branch metric m[] and each state at each node level by its state metric M[], the state metrics at node level K are obtained from the state metrics at the level K - 1by adding to each state metric at level K - 1 the branch metrics that connect it to states at the Kth level, and for each state at level K preserving only the largest sum that arrives to it. If additionally at each level we delete all branches other than the one that produces this maximum, there will remain only one path through the trellis leading from the origin to each state at the *K*th level, which is the most probable path reaching it from the origin. In typical (but not all) applications, both the initial state (origin) and the final state are fixed to be S_0 and thus the algorithm produces the most probable path through the trellis both initiating and ending at S_0 .

2. APPLICATIONS

Numerous applications of this algorithm have appeared since the 1960s or so. The following list represents the most prominent in approximate chronological order:

- 1. Decoders for convolutional codes on various wireless channels
- 2. Maximum-likelihood sequence estimation (MLSE) demodulators for intersymbol interference and multipath fading channels
- 3. Decoders for recorded data
- 4. Optical character recognition
- 5. Voice recognition
- 6. DNA sequence alignment

We shall describe each in the order indicated above.

2.1. Convolutional Codes

The earliest application, for which the algorithm was originally proposed in 1967, was for the maximum likelihood decoding of convolutionally coded digital sequences transmitted over a noisy channel. Currently the algorithm forms an integral part of the majority of wireless telecommunication systems, both involving satellite and terrestrial mobile transmission. The convolutional encoder and channel combination, as shown in Fig. 3, gives rise directly to the Markov graph representation. In the simplest case, one bit at a time enters the L-stage shift register and the n linear combiners, each of which is a modulo-2 adder of the contents of some subset of the Lshift register stages, generate n binary symbols. These are serially transmitted, for example, as binary amplitude (x = +1 or -1) modulation of a carrier signal. At the receiver, the demodulator generates an output y, which is either a real number or the result of quantizing the latter to one of a finite set of values. The conditional densities p(y/x) of the channel outputs are assumed to be mutually independent, corresponding to a "memoryless channel." A commonly treated example of such is the additive white Gaussian noise (AWGN) channel for which each y is the sum of the encoded symbol x and a Gaussian random noise variable, with all noise variables mutually independent. This channel model is closely approximated by satellite and space communication applications and, with appropriate caution, it can also be applied to terrestrial communication design.



Figure 3. Convolutionally encoded memoryless channel.

The communication system model just described gives rise naturally to the Markov graph representation. The 2^{L} states correspond to the states of the contents of the L-stage register. Thus S_0 corresponds to the contents being all zeros, S_1 to the first stage containing a one and all the rest zeros, and so on. Since only one input bit changes each time, each state has only two branches both exiting and entering it, each from two other states. For the exiting branches, one corresponds to a zero entering the register and the other to a one. Figure 1 could be used to represent a two-stage encoder, where the state indices are the decimal equivalents of the binary register contents. It is generally assumed that all input bits are equally likely to be a zero or a one, so the state transition probabilities, $P(S_i \rightarrow S_i) = \frac{1}{2}$ for each branch. Hence the first term of the branch metric m can be omitted since it is the same for each branch. As for the second term of m, the conditional probability density $p(y/S_i \rightarrow S_i) = p(y/x)$, where *x* is an *n*-dimensional binary vector generated by the *n* modulo-2 adders for each new input bit, which corresponds to a state transition, while *y* is the random vector corresponding to the *n* noisecorrupted outputs for the *n* channel inputs represented by the vector x. For the AWGN, $\ln p(y/x)$ is proportional to the inner product of the two vectors *x* and *y*.

The convolutional encoder and its Markov graph just described represent a rate 1/n code, since each input bit generates n output symbols. To generalize to any rational rate m/n < 1, m input bits enter each time and the register shifts in blocks of m. The Markov graph changes only in having each state connected to 2^m other states. Another generalization is to map each binary vector x, not into a vector of n binary values, +1 or -1, but into a constellation of points in two or more dimensions. An often employed case is quadrature amplitude modulation (QAM). For example, for n = 4, 16points may be mapped into a two-dimensional (2D) grid and the value in each dimension modulates the amplitude of one of the two quadrature components of the sinusoidal carrier. Here *x* is the 2D vector representing one of the 16 modulating values and y is the corresponding demodulated channel output. Multiple generalizations of this approach abound in the literature and in real applications. In most cases this multidimensional approach is used to conserve bandwidth at the cost of a higher channel signal-to-noise requirement.

An interesting footnote on this first application is that the Viterbi algorithm was proposed not so much to develop an efficient maximum-likelihood decoder for a convolutional code, but primarily to establish bounds on its error correcting performance.

2.2. MLSE Demodulators for Intersymbol Interference and Multipath Fading Channels

In the previous application, the convolution operation is employed in order to introduce redundancy for the purpose of increasing transmission reliability. But convolution also occurs naturally in physical channels whose bandwidth constraints linearly distort the transmitted digital signal. Treating the channel as a linear filter, it is well known that the output signal is the convolution of the input signal and the filter's impulse response. A discrete model of the combination of signal waveform, channel effects, and receiver filtering is shown in Fig. 4. This combination produces, after sampling at the symbol

rate, the discrete convolution $x_k = \sum_{j=0} h_j u_{k-j}$, where the variables u are the input bit sequence, generally taken to be binary (+A or -A). The h_j terms are called the *intersymbol interference* coefficients, since except for the h_0 term, all the other terms of the sum represent interference by preceding symbols on the given symbol. To the output of the discrete filter x_k must be added noise variables n_k to account for the channel noise. While generally these noise variables are not mutually independent, they can be made so by employing at the receiver a so-called whitened matched filter prior to sampling.

A related application with a very similar model is that of multipath fading channels. Here the taps represent multiple delays in a multipath channel. Their relative spacings may be less than one symbol period, in which case each symbol of the input sequence must be repeated several times. Most important, because of random variations in propagation, the multipath coefficients hare now random variables, so the conditional densities of y depend on the statistics of both the additive noise and the multiplicative random coefficients.

Comparing Fig. 4 with Fig. 3, we note that the principal difference is that modulo-2 addition is replaced by real addition and there is one rather than n outputs for each branch. Otherwise, the same Markov graph applies, with two branches emanating from each state when the input sequence is binary (+A or -A). Generation of the branch metrics, $\ln p(y/S_j \rightarrow S_i)$, is slightly more complex because besides depending on the noise, the y variables depend on a linear combination of the h variables, with their signs



Figure 4. Intersymbol interference or multipath fading channel model.

2818 VITERBI ALGORITHM

determined by the register contents involved in the branch transition.

2.3. Partial-Response Maximum-Likelihood Decoders for Recorded Data

A filter model for the magnetic recording medium and reading process is very similar to the intersymbol interference model. The simplest version, known as the "partial response" channel results, when sampled at the symbol rate, in an output depending on just the difference between two input symbols, $x_k = u_k - u_{k-2}$, with additive noise samples also being mutually independent. Note that since inputs u are +1 or -1, the outputs x (prior to adding noise) are ternary, +2, 0, or -2. This then reduces to the model of Fig. 4 with just two nonzero taps, $h_0 = +1$ and $h_2 = -1$. Often this is described by the polynomial whose coefficients are the tap values; in this case $h(D) = (1 - D^2)$. This can be modeled by a twostage shift register that gives rise to a four-state Markov graph, as in Figs. 1 and 2. But actually, a simpler model can be used based on the fact that all outputs for which k is odd depend only on odd-indexed inputs, and similarly for even. Thus a two-state Markov graph suffices for each of the odd and even subsets, as shown in Fig. 5. When the recording density is increased, the simple partialresponse channel model is replaced by a longer shift register. A generally accepted model has tap coefficients represented by the polynomial $h(D) = (1-D)(1+D)^N$, where N > 1. For example, for the case of N = 2, known as "extended partial response," an eight-state Markov graph applies.

The Markov nature of the preceding three applications is obvious from the system model. For the next three, it is not as obvious and often it is only a tentative model of the phenomenon derived from observations. In such cases the term *hidden Markov model* (HMM) is used. Such models are often empirically derived based on experience in the given discipline. Since background in each field is a prerequisite for full understanding, we give only a superficial description of each of the following.

2.4. Optical Character Recognition

The algorithm has been applied to the automatic character recognition of hand-printed text. For many decades, statistical analysis of English (and other languages) has led to a Markov model whose states are the single letters, digrams, trigrams, or generally N-grams, of English text, as measured by their relative frequency in numerous published texts. Thus given 27 letters, including space

as a letter, a Markov graph of 27^N states can be created. Unlike the previous applications, here the branch metric

$$\begin{split} m[y(k); S(k-1), S(k)] \\ &= \ln\{\Pr[S(k-1) \to S(k)] + \ln p[y(k)/S(k-1) \to S(k)]\} \end{split}$$

depends as much on the first term as on the second. The first term, of course, is determined, as just noted, from the predetermined N-gram transition relative frequencies. (The transitions will involve just a change of a single letter as, e.g., transition from THA to HAT.) As for the observables y(k), these may be as simple as the character recognizer's preliminary estimate (without benefit of the Markov character) to measurements on a grid of points, possibly including gray levels, which will result in better performance. In any case, the conditional density of the measurement values given the letter causing the branch transition must also be predetermined to implement the second term.

2.5. Voice Recognition

An increasingly popular application has been to voice recognition for automated dialing or response systems. The situation is very similar to character recognition, except that the language text characters are replaced by *phonemes*, which are voiced fragments of speech. So the states may be the phoneme N-grams and the first term of the branch metrics are the predetermined relative frequencies of transitions between phonemes. The second term then depends on measurements of the recorded voice phoneme and its conditional probability relative to the actual phoneme causing the given transition.

2.6. DNA Sequence Analysis

The most recent and most surprising application has been to the alignment of strands of DNA sequences involved in mapping the genome. DNA sequences consist of four types of nucleotides labeled A, C, G, and T. In addition, in analyzing similarities among sequences, one must accept the possibility of insertions and deletions. Thus the states of the hypothetical hidden Markov model each involve nucleotides, insertions, or deletions or some combination thereof. The Markov state-transition probabilities are initially assigned arbitrarily or based on previous experience. After the maximum-likelihood alignments are found by means of the algorithm, the relative frequencies of the state transitions are counted and used as the transition probabilities for a second iteration of the algorithm. This process may continue



Figure 5. Markov graph and Trellis diagram for even (odd) partial-response states.

States in \square , branches labelled with *x* values

through several iterations until the measured relative frequencies of a given iteration provide a good match to those of the hypothesized model from the previous iteration. In this way the accuracy and reliability of the HMM is refined along with the maximum-likelihood alignments.

2.7. Other Applications

Given the pervasiveness of Markov models for a variety of fields, there have been numerous other applications of the Viterbi algorithm throughout the engineering and scientific literature, and there will likely continue to be more. The list is too lengthy and some applications are too obscure to identify. We have provided here the six most often cited.

BIOGRAPHY

Dr. Andrew Viterbi is a cofounder, retired vice chairman, and chief technical officer of QUALCOMM Incorporated. He spent equal portions of his career in industry, having also cofounded a previous company, and in academia as a professor in the Schools of Engineering first at UCLA and then at UCSD, at which he is now professor emeritus. He is currently president of the Viterbi Group, a technical advisory and investment company.

His principal research contribution, the Viterbi algorithm, is used in most digital cellular phones and digital satellite receivers, as well as in such diverse fields as magnetic recording, voice recognition, and DNA sequence analysis. In recent years he has concentrated his efforts on establishing CDMA as the multiple access technology of choice for cellular telephony and wireless data communication.

Dr. Viterbi has received numerous honors both in the United States and internationally. Among these are four honorary doctorates, from the Universities of Waterloo, Rome, Technion, and Notre Dame, as well as memberships in the National Academy of Engineering, the National Academy of Sciences, and the American Academy of Arts and Sciences. He has received the Marconi International Fellowship Award, the IEEE Alexander Graham Bell and Claude Shannon Awards, the NEC C&C Award, the Eduard Rhein Award, and the Christopher Columbus Medal.

VOCODERS

ANDREAS SPANIAS Arizona State University Tempe, Arizona

1. INTRODUCTION

Vocoder is a term that is formed from the concatination of the words *voice* and *coder*. Although originally it became associated with a specific class of analysis/synthesis systems for speech bandwidth reduction, such as the channel vocoder, it is now used for a wider class of algorithms for the compression of bit rate of speech signals. Strictly speaking, *vocoders* as opposed to *waveform coders* represent a subcategory of speech coders that make heavy use of speech spectral properties and rely on a source system model for the representation and parametrization of speech. More recently the term *vocoder* has been used more loosely, and essentially it became associated with the speech coding algorithms that are used in cellular phones, streaming speech applications, Internet telephony, secure communications, digital answering machines, portable digital voice recorders, and other applications.

Vocoder algorithms are embedded in several international standards formed for telephony and multimedia applications. The standardization section of the International Telecommunications Union (ITU), formerly CCITT, has been developing compatibility standards for telephony and more recently for Internet and multimedia applications. Other standardization committees, such as the European Telecommunications Standards Institute (ETSI) and the International Standards Organization (ISO), have also drafted requirements for speech (GSM) and audio/video coding (MPEG) standards. In addition to these organizations there are also committees forming standards for private or government applications such as secure telephony, satellite communications, and emergency radio applications. Standard specifications have driven much of the research and development in the speech coding area and several speech and audio coding algorithms have been developed and eventually adopted in international standards. A series of competing speech signal models based on linear predictive coding (LPC) [1-7] and transform-domain analysis-synthesis [8-10] have been proposed since the mid-1980s. On the other hand, high-end audio coding relied heavily on sub-band and transform coding algorithms that use psychoacoustic signal models [11,12].

Speech coding for low-rate applications involves parametric representation of speech using analysis synthesis systems. Analysis can be open-loop or closedloop. In closed-loop analysis, also called analysis-bysynthesis, the parameters are extracted and encoded by minimizing the perceptually weighted difference between the original and reconstructed speech. Speech coding algorithms are evaluated based on speech quality, algorithm complexity, delay, and robustness to channel and background noise. Moreover in network applications coders must perform reasonably well with nonspeech signals such as DTMF tones, voiceband data, and modem tones. Standardization of candidate speech coding algorithms involves evaluation of speech quality using subjective measures such as the Mean Opinion Score (MOS), which involves rating speech according to a 5-level quality scale, as shown in Table 1.

A MOS of 4–4.5 is associated with network or toll quality (wireline telephone grade), and scores between 3.5 and 4 imply communications quality (cellular grade). The simplest coder that achieves toll quality is the 64 kbps (kilobits per second) ITU G.711 pulse code modulation (PCM), which has a MOS of 4.3. Several of the new general-purpose algorithms, such as the 8 kbps ITU G.729 [13] and 6.3 kbps G.723.1 [38], also achieve toll quality. Algorithms

Table 1. The MOS Scale

MOS	Subjective Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

for the cellular standard such as the *TIA* IS54 [14], the fullrate ETSI GSM 6.10 [15] achieve communications quality, and the old Federal Standard 1015 (LPC-10e) [16] is associated with synthetic quality. More recent algorithms for cellular standards have near-toll quality performance. Such algorithms include the adaptive multirate (AMR) coder for the GSM system [42,43] and the selectable mode vocoder (SMV) [52] for use in wideband CDMA applications. The new frontier in vocoder standardization targets toll quality at 4 kbps. In fact, ITU is currently evaluating several algorithms [45–51] that aim for toll quality at 4 kbps.

In this article, we will focus on speech coding algorithms. Section 2 presents an introduction to speech properties and the opportunities for bit rate reduction. Section 3 discusses source system representations of speech and the use of linear prediction in speech coding. Section 4 presents open-loop algorithms and the classical LPC-10 algorithm. Section 5 presents closed-loop LP algorithms, and Section 6 focuses on code-excited linear prediction (CELP) and three generations of standardized algorithms based on CELP. Section 7 concludes with a summary of this article.

2. THE SPEECH SPECTRAL PROPERTIES AND VOCODERS

Before we begin our presentation of vocoder algorithms, we discuss some of the important speech spectral properties that provide opportunities for speech parametrization and compression. First, in digital telephony applications speech is typically band-limited to 3.2 kHz and sampled at 8 kHz. The bandwidth of 3.2 kHz preserves both speech intelligibility and the speaker identity. Speech is a nonstationary random signal and is considered as quasistationary only over short segments, typically 5-20 ms. Speech can generally be classified as voiced (e.g., /a/, /i/), unvoiced (e.g., /sh/), or mixed. We note that this course classification into voiced/unvoiced/mixed segments is adequate only for coding applications. Speech recognition and voice synthesis algorithms used a finer and much more precise phonemic classification. Timedomain plots for sample voiced and unvoiced segments are shown in Fig. 1. A segment of voiced speech from an uttered steady vowel is quasiperiodic in the time domain and harmonically structured in the frequency domain. Unvoiced speech is randomlike and broadband. In addition, the energy of voiced segments is generally higher than the energy of unvoiced segments.

The short-time spectrum of *voiced* speech is characterized by its fine and formant structure. The fine harmonic structure is due to the periodicity of voiced speech and



Figure 1. Voiced (top) and unvoiced (bottom) waveforms and their short-time spectra.

is attributed to the activity of the vibrating vocal chords. The *formant* structure (spectral envelope) is due to the interaction of the source and the vocal tract. The spectral envelope shown in Fig. 2 "fits" the short-time spectrum of voiced speech and is associated with the transfer characteristics of the vocal tract and the spectral tilt (6 dB/octave) due to the glottal pulse. The spectral envelope is characterized by a set of peaks called *formants*. The formants are the resonant modes of the vocal tract. For the average vocal tract there are three to five formants below 5 kHz.

The amplitudes and locations of the first three formants, which usually occur below 3 kHz, are quite important in both speech synthesis and perception. Higher formants are also important for wideband and unvoiced speech representations. The properties of speech are related to the physical speech production system as follows. Voiced speech is produced by exciting the vocal tract with quasiperiodic glottal air pulses generated by the vibrating vocal chords. The frequency of the periodic pulses is referred to as the *fundamental frequency* or "pitch." Unvoiced speech is produced by forcing air through a constriction in the vocal tract. Nasal sounds (e.g., /n/) are due to the acoustical coupling of the nasal tract to the vocal tract, and plosive sounds (e.g., /p/) are produced by abruptly releasing air pressure that was built up behind a closure in the tract.

3. SOURCE SYSTEM MODELS AND SHORT-TERM LINEAR PREDICTION

Speech is produced by the interaction of the vocal tract with the vocal chords in the glottis. Engineering models (Fig. 3) for speech production typically model the vocal tract as a time-varying digital filter excited by quasiperiodic waveform when speech is voiced (e.g., as in steady vowels) and random waveforms for unvoiced speech (e.g., as in consonants). The vocal tract filter is estimated using linear prediction (LP) algorithms [17,18].

Linear prediction algorithms are part of several speech coding standards, including ADPCM systems [19–21], open-loop linear predictive coders [16], and *code-excited linear prediction* (CELP) algorithms and other analysisby-synthesis linear predictive coders [21–26]. In linear prediction the most recent sample of speech is predicted by a linear combination of past samples. This is done using a finite-length impulse response (FIR) filter (Fig. 4) whose





Figure 3. Engineering model for speech synthesis.



Figure 4. Linear prediction analysis.

output e(n) is minimized. The output of the linear predictor analysis filter is called the linear prediction residual.

The LP coefficients are chosen to minimize the mean square of the LP residual

$$\varepsilon = E[e^2(n)] \tag{1}$$

where the error is the difference of the current speech sample and a linear combination of past samples:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$
 (2)

Because only short-term delays are considered in (2), the linear predictor in Fig. 4 is also known as a *short-term linear predictor*. The inverse of the LP analysis filter is an all-pole filter called the *LP synthesis filter*. The frequency response associated with the short-term synthesis filter captures the *formant* structure of the short-term speech spectrum. The all-pole filter or vocal tract transfer function is given by

$$H(z) = \frac{g}{1 - \sum_{k=1}^{M} a_k z^{-k}}$$
(3)

The minimization of ε yields a set of equations involving autocorrelations $[r_{ss}(m)]$ of speech and the LP coefficients

Figure 2. (a) Spectral envelope and (b) formants.

are found by inverting an autocorrelation matrix using the Levinson–Durbin algorithm [24].

$$\begin{bmatrix} a_{1} \\ a_{2} \\ a_{3} \\ \vdots \\ a_{M} \end{bmatrix} = \begin{bmatrix} r_{xx}(0) & r_{ss}(-1) & r_{ss}(-2) & \cdots & r_{xx}(1-M) \\ r_{ss}(1) & r_{ss}(0) & r_{ss}(-1) & \cdots & r_{ss}(2-M) \\ r_{ss}(2) & r_{ss}(1) & r_{ss}(0) & \cdots & r_{ss}(3-M) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{ss}(M-1) & r_{ss}(M-2) & r_{ss}(M-3) & \cdots & r_{ss}(0) \end{bmatrix}^{-1} \\ \times \begin{bmatrix} r_{ss}(1) \\ r_{ss}(2) \\ r_{ss}(3) \\ \vdots \\ r_{ss}(M) \end{bmatrix}$$
(4)

There are many efficient algorithms [17,18] for inverting this matrix including algorithms tailored to work well with finite-precision arithmetic [33]. Preconditioning of the speech and autocorrelation data using tapered windows improves the numerical behavior of these algorithms. In addition, bandwidth expansion or scaling of the LP coefficients is very typical in LPC as it reduces distortion during synthesis.

In short-term LP the analysis window is typically 20 ms long. In order to avoid transient effects from large changes in the LP parameters from one frame to the next, the frame is usually divided into subframes (typically 5 ms long), and subframe parameters are obtained by linear interpolation. The direct form LP coefficients a_k are not adequate for quantization and transformed coefficients are typically used in quantization tables. The reflection or lattice prediction coefficients are a byproduct of the Levinson recursion and have better quantization properties than do direct-form coefficients. Some of the early standards such as the LPC-10 [16] and the IS54 VSELP [14] encode reflection coefficients for the vocal tract. Transformation of the reflection coefficients can also lead to a set of parameters that are less sensitive to quantization. In particular, the log area ratios (LARs) and the inverse sine transformation have been used in the early GSM 6.10 algorithm [15] and in the skyphone standard [22]. Most recent LPC-related cellular standards [23-25] quantize line spectrum pairs (LSPs). The main advantage of the LSPs is that they relate directly to frequencydomain information and hence they can be encoded using perceptual criteria. In most recent toll-quality standards such as the wideband CDMA SMV [52], the GSM adaptive multirate vocoder [42,43], the ITU G.723.1 [38], and the

2822 VOCODERS

ITU G.729 [13] the LSPs are encoded using split-vector quantization.

3.1. Linear Prediction and ADPCM

One of the simplest scalar quantization schemes that uses short-term LP is the adaptive differential pulse code modulation (ADPCM) coder [19,20]. ADPCM algorithms encode the difference between the current and the predicted speech samples. The prediction parameters are obtained by backward estimation, namely, from quantized data, using a gradient algorithm. The ADPCM 32-kbps algorithm in the ITU G.726 standard (formerly known as CCITT G.721) uses a pole-zero adaptive predictor (Fig. 5). ITU G.726 also accommodates 16, 24, and 40 kbps with individually optimized quantizers. The ITU G.727 has embedded quantizers and was developed for packet network applications. Because of the embedded quantizers, G.727 has the capability to switch easily to lower rates in network congestion situations by dropping bits. The MOS for 32 kbps G.726 is 4.1, and complexity is estimated to be 2 million instructions per second (Mips) on special-purpose chips and about 10 Mips on generic fixed-point DSP chips.

The International Mobile Satellite B (INMARSAT-B) standard [27] uses also a 10MIPS ADPCM coder with a long-term predictor (discussed later) in addition to short-term LP. The INMARSAT-B algorithm operates at 12.8 and 9.6 kbps.



Figure 5. The ITU G.726 ADPCM encoder.

4. SPEECH ANALYSIS-SYNTHESIS USING LINEAR PREDICTION

Unlike waveform ADPCM coders that use LP only for differential quantization, this class of algorithms use LP to represent the vocal tract and make explicit use of the synthesis model shown in Fig. 3. Our discussion of linear prediction is divided in two categories: openloop analysis-synthesis LP and closed-loop analysis-bysynthesis linear prediction. In the following, we describe two open-loop algorithms, the LPC-10 and the mixedexcitation LPC. Unless otherwise stated, the input to all coders discussed in this section is speech sampled at 8 kHz.

4.1. Open-Loop Linear Prediction

This section describes source system algorithms that use open-loop analysis to determine the excitation sequence. Open-loop linear predictive vocoders are essentially the first-generation LP vocoders. The DOD LPC-10 is a good example of an algorithm that uses open-loop analysis. In 1976 a consortium established by the U.S. Department of Defense (DoD) recommended an LPC algorithm for secure communications at 2.4 kbps, (Fig. 6). The algorithm, known as the LPC-10, eventually became the original Federal Standard FS-1015 [16,28,29]. The LPC-10 uses a 10th-order predictor to estimate the vocal tract parameters. Segmentation and frame processing in LPC-10 depend on voicing. Pitch information is estimated using the average magnitude difference function (AMDF) [16]. Voicing is estimated using energy measurements, zerocrossing measurements, and the maximum to minimum ratio of the AMDF. The excitation signal for voiced speech in the LPC-10 consists of a sequence that resembles a sampled glottal pulse. This sequence is defined in the standard [16] and periodicity is created by a pitch-synchronous pulse repetition process. The LPC-10 produces synthetic speech with a MOS of 2.3. Complexity is estimated at 5-7 Mips.

4.2. Mixed-Excitation Linear Prediction

In 1996 the U.S. government standardized a new 2.4kbps algorithm called *mixed-excitation LP* (MELP) [7]. The development of mixed-excitation models in LPC was motivated largely by voicing errors in LPC-10 and also by the inadequacy of the two-state excitation model



Figure 6. The LPC-10 encoder.



in cases of voicing transitions (mixed voiced-unvoiced frames) [7,30]. This problem is solved using a mixed excitation scheme that combines the lowband impulse train (buzz) and highband noise (Fig. 7). The excitation shaping is done using first-order FIR filters $H_1(z)$ and $H_2(z)$ with time-varying parameters. The mixed source model also uses (selectively) pulse position jitter for the synthesis of weakly periodic or aperiodic voiced speech. An adaptive pole-zero spectral enhancer is used to boost the formant frequencies. Finally a dispersion filter is used after the LP synthesis filter to improve the matching of natural and synthetic speech away from the formants. The 2.4-kbps MELP was based on a 22.5-ms frame, and the algorithmic delay was estimated to be 122.5 ms. An integer pitch estimate is obtained open-loop by searching autocorrelation statistics followed by a fractional pitch refinement process. The LP parameters are obtained using the Levinson-Durbin algorithm and vector quantized as LSPs. MELP outperforms the LPC-10 with an estimated MOS of 3.2 but with higher complexity estimated at 40 Mips.

5. ALGORITHMS BASED ON ANALYSIS-BY-SYNTHESIS LINEAR PREDICTION

We describe here several speech coding standards based on a class of modern source-system coders where system parameters are determined by linear prediction and the excitation sequence is determined by closedloop or analysis-by-synthesis optimization (Fig. 8). The optimization process determines an excitation sequence that minimizes the weighted difference between the input speech and synthesis speech [1-3]. Strictly speaking, this class of speech coders are not called vocoders but instead they are called *hybrid* coders. This is because closed-loop LP combines the spectral modeling properties of vocoders with the waveform-matching features of waveform coders.



The system consists of a short-term LP synthesis filter, a long-term LP synthesis filter for the pitch (fine) structure of speech, a perceptual weighting filter W(z) that shapes the error such that quantization noise is masked by the high-energy formants, and the excitation generator. The three most common excitation models for analysis-by-synthesis LPC are the multipulse model [2,3], the regular pulse excitation model [5], and the vector or code excitation model [1]. These excitation models are described in the context of standardized algorithms.

5.1. Long-Term Prediction (LTP)

Almost all *analysis-by-synthesis* LP algorithms include long-term prediction in addition to short-term prediction. Long term prediction, as opposed to the short-term prediction, is a process that captures the long-term correlation in the speech signal. The LTP provides a mechanism for representing the periodicity in speech and as such it represents the *fine* harmonic structure in the short-term speech spectrum. The LTP requires estimation of two parameters: a delay a_{\leftrightarrow} and a parameter a_{\leftrightarrow} . For strongly voiced segments the delay is usually an integer that approximates the pitch period. A transfer function of a simple LTP synthesis filter is given below; more complex LTP filters involve multiple parameters and noninteger (fractional) delays [26]:

$$H_{\tau}(z) = \frac{1}{1 - a_{\tau} z^{-\tau}}$$
(5)

The LTP can be implemented as open loop or closed loop. The open-loop LTP parameters are typically obtained by searching the autocorrelation sequence over 128 integer delays (20–147 for speech sampled at 8 kHz). The gain is simply obtained by $a_{\leftrightarrow} = r_{ss}(\leftrightarrow)/r_{ss}(0)$. Closed-loop LTP searches produce improved speech quality at the expense of additional complexity. In closed-loop LTP search the



Figure 8. MPLP for the Skyphone standard.

signal is synthesized for a range of candidate LTP lags and the lag that produces the best waveform matching is chosen. Because of the intensive computations in fullsearch closed-loop LTP, more recent algorithms use openloop LTP to establish an initial LTP lag, which is then refined using closed-loop search around the neighborhood of the initial estimate. In addition, to reduce complexity further LTP searches are in some cases carried every other subframe.

5.2. Multipulse Excited Linear Prediction

A 9.6-kbps multipulse excited linear prediction (MPLP) algorithm is used in *skyphone* airline applications [22]. The MPLP algorithm forms an excitation sequence that consists of multiple nonuniformly spaced pulses (Fig. 8). During analysis both the amplitude and locations of the pulses are determined (sequentially) one pulse at a time such that the weighted mean-square error is minimized. The MPLP algorithm typically uses 4–6 pulses every 5 ms [2,3]. The weighting filter is given by

$$W(z) = \frac{1 + \sum_{i=1}^{p} \gamma_1^i a_i z^{-i}}{1 + \sum_{i=1}^{p} \gamma_2^i a_i z^{-i}} \quad 0 < \gamma_2 < \gamma_1 < 1$$
(6)

The role of W(z) is to deemphasize the error energy in the formant regions. This deemphasis strategy is based on the fact that in the formant regions quantization noise is partially masked by speech. Excitation coding in the MPLP algorithm is more expensive than in the classical linear predictive vocoder because MPLP encodes both the amplitudes and the locations of the pulses. The British Telecom International skyphone MPLP algorithm accommodates passenger communications in aircraft. The algorithm incorporates both short- and longterm prediction. The LP analysis window is updated every 20 ms and the LTP parameters are obtained using openloop analysis. The MOS for the skyphone algorithm is 3.4.

5.3. The Regular Pulse Excitation (RPE) Algorithm

RPE coders also employ an excitation sequence which consists of multiple pulses. The basic difference of the RPE algorithm from the MPLP algorithm is that the pulses in the RPE coder are uniformly spaced and therefore their positions are determined by specifying the location of the first pulse within the frame and the spacing between nonzero pulses. The analysis-by-synthesis optimization in RPE algorithms represents the LP residual by a regular pulse sequence that is determined by weighted error minimization [5]. A 13-kbps coding scheme that uses RPE with long-term prediction (LTP) was adopted in 1990 for the full-rate ETSI GSM [22] Pan-European digital cellular standard. The performance of the GSM codec in terms of MOS was reported to be between 3.47 (min) and 3.9 (max), and its complexity is 5 to 6 Mips.

6. CODE-EXCITED LINEAR PREDICTION (CELP) ALGORITHMS

The vector or code-excited linear prediction (CELP) algorithm [1] (Fig. 9) encodes the excitation using vector quantization. The codebook used in a CELP coder contains vector excitation, and in each subframe a vector is chosen using an analysis-by-synthesis process. The "optimum" vector is selected such that the perceptually weighted MSE is minimized. A scaled excitation vector is filtered by the long- and short-term synthesis filters.

This category of analysis-by-synthesis LPC algorithms proved to be the most successful in terms of standards and applications. We describe in the following CELP algorithms as used in a variety of standards. We divide these algorithms in three categories that are also consistent with the chronology of their development: firstgeneration CELP (1986–1992), second-generation CELP (1993–1998), and third-generation CELP (1999–present).

6.1. First-Generation CELP Coders

Although the initial development of some of these algorithms was funded in part by the U.S. Department of Defense (DoD), the key driving force behind research and development of CELP coders was the demand for vocoders for the first generation digital cellular phones. Many of the first generation CELP algorithms were developed between 1986 and 1992 and work at bit rates of 4.8–16 kbps. These are generally high complexity algorithms and nontoll quality. These algorithms include, the FS-1016 CELP, the IS54 VSELP, the IS96 QCELP, and the G.728 LD-CELP.

A 4.8-kbps CELP algorithm is used by the Department of Defense for possible use in the third-generation secure telephone unit (STU-III) [31,32]. This algorithm



Figure 9. The analysis-by-synthesis CELP algorithm.

is described in the Federal Standard 1016 (FS-1016) and was jointly developed by the DoD and the Bell Labs. Speech in the FS-1016 CELP is sampled at 8 kHz and segmented in frames of 30 ms, and each frame is segmented in subframes of 7.5 ms. The excitation in this CELP is formed by combining vectors from an adaptive (LTP) and a stochastic codebook. The excitation vectors are selected in every subframe and the codebooks are searched sequentially starting with the adaptive codebook. The term *adaptive codebook* is used because the backward estimated LTP lag search can be viewed as an adaptive codebook search where the codebook is defined by previous excitation sequences (LTP state), and the lag τ determines the vector index. The adaptive codebook contains the history of past excitation signals and the LTP lag search is carried over 128 integer (20-147) and 128 noninteger delays. A subset of lags is searched in even subframes to reduce the computational complexity. The stochastic codebook contains 512 sparse and overlapping codevectors. Each codevector consists of sixty samples and each sample is ternary valued (1, 0, -1) to allow for fast convolution. Ten short-term prediction parameters are encoded as LSPs on a frame-by-frame basis. Sub-frame LSPs are obtained by linear interpolation. The computational complexity of the FS1016 CELP was estimated at 16 MPS, and a MOS score of 3.2 has been reported.

Another standardized CELP is the IS54 VSELP, which uses highly structured codebooks that are tailored for reduced computational complexity and increased robustness to channel errors. VSELP excitation is derived by combining excitation vectors from three codebooks, namely, an adaptive codebook and two highly structured stochastic codebooks, (Fig. 10). The 128 Forty-sample vectors in each stochastic codebook are formed by linearly combining seven basis vectors. The weights used for the basis vectors are allowed to take the values of one or minus one. Hence the effect of changing one bit in the codeword, possibly due to a channel error, is not minimal since the *codevectors* corresponding to adjacent (gray codewise) codewords are different only by one basis vector. The search of the codebook is also greatly simplified because the response of the short-term synthesis filter, to codevectors from the stochastic codebook, can be formed

by combining filtered basis vectors. The complexity of the 8-kbps VSELP was reported to be around 13.5 MPS, and the MOS reported was 3.45. The vector-sum-excited linear prediction (VSELP) algorithm [6] and its variants are embedded in three digital cellular standards: the 81-kbps TIA IS54 [14], the 6.3-kbps Japanese standard [33], and the 5.6-kbps half-rate GSM [34].

One problem in network applications of speech coding is that coding gain is achieved at the expense of coding delay. The one-way delay is basically the time elapsed from the instant a speech sample arrived at the encoder to the instant that this sample appears at the output of the decoder. This definition of one-way delay does not include channel- or modem-related delays. Roughly speaking, the one-way delay is generally between two and four frames. The ITU G.728 low-delay CELP coder [35,36] achieves low one-way delay by short frames, backward-adaptive predictor, and short excitation vectors (5 samples). In backward-adaptive prediction, the LP parameters are determined by operating on previously quantized speech samples that are also available at the decoder, (Fig. 11). The LD-CELP algorithm does not utilize LTP. Instead, the order of the short-term predictor is increased to 50 to compensate for the lack of a pitch loop.

The frame size in LD-CELP is 2.5 ms, and the subframes are 0.625 ms long. The parameters of the 50thorder predictor are updated every 2.5 ms. The perceptual weighting filter is based on 10th-order LP operating directly on unquantized speech and is updated every 2.5 ms. In order to limit the buffering delay in LD-CELP only 0.625 ms of speech data are buffered at a time. LD-CELP utilizes adaptive short- and long-term postfilters to emphasize the pitch and formant structures of speech. The one-way delay of the LD-CELP is less than $2\ \mathrm{ms}$ and MOSs as high as 3.93 and 4.1 were obtained. The speech quality of the LD-CELP was judged to be equivalent or better than the G.726 standard even after three asynchronous tandem encodings. The coder was also shown to be capable of handling voiceband modem signals at rates as high as 2400 baud (provided that perceptual weighting is not used). The coder complexity and memory requirements were found to be 10.6 MPS and 12.4 kB for the encoder and 8.06 MPS and 13.8 kB for the decoder.



Figure 10. The IS54 VSELP algorithm.



Figure 11. The G.728 low-delay CELP algorithm.

Figure 12. The IS96 QCELP decoder.

6.1.1. Variable-Rate CELP Algorithms for CDMA Applications. The IS96 QCELP [37], (Fig. 12) is a variable bit rate algorithm and is part of the code-division multiple-access (CDMA) standard for cellular communications. The bit rate is variable, with four rates supported" 9.6, 4.8, 2.4, and 1.2 kbps. The rate is determined by speech activity. Rate changes can also be activated upon command from the network. The short-term LP parameters are encoded as LSPs.

Lower rates are achieved by allocating fewer bits to LP parameters and by reducing the number of updates of the LTP and random codebook parameters. At 1.2 kbps (rate $\frac{1}{8}$) the algorithm essentially encodes comfort noise. The MOS for QCELP at 9.6 kbps is 3.33, and the complexity is estimated to be around 15 MPS.

6.2. Second-Generation Near-Toll-Quality CELP Algorithms

The second-generation CELP algorithms were developed between 1993 and 1998 for applications in secondgeneration cellular phones, PC Internet streaming applications, Voice over Internet Protocol (VoIP), and secure communications. These are also high-complexity algorithms that deliver near-toll-quality speech, with some of the most successful ones certified for toll quality. The speech quality enhancement with these coders is because of the improvement in the coding of excitation, which is done by algebraic codebooks [13], as well as due to the coding of the LPC parameters in using line spectral frequencies and perceptually optimized split-vector quantization. Furthermore interpolative techniques with the LTP (relaxed CELP) [4] also contribute to improvements. Algorithms in this category include the G.729 ACELP, the G.723.1 coder, the GSM EFR, the IS127 RCELP.

6.2.1. CELP Algorithms with Algebraic Codebooks. A low-delay 8 kbps conjugate structure algebraic CELP (CS-ACELP) algorithm has been adopted as ITU recommendation G.729 [13]. The G.729 is designed for both wireless and multimedia network applications. CS-ACELP is a lowdelay algorithm with a frame size of 10 ms, a lookahead of 5ms, and a total algorithmic delay of 15 ms. The algorithm is based on an analysis-by-synthesis CELP scheme and uses two codebooks for excitation modeling. The shortterm prediction parameters are obtained every 10 ms and vector-quantized as LSPs. The algorithm uses an algebraically structured fixed codebook that does not require storage. Each 40-sample (5-ms) codevector contains four nonzero, binary-valued (-1, 1) pulses. Pulses are interleaved and position-encoded, which allows efficient search. Gains for the fixed and adaptive codebooks are jointly vector-quantized in a two-stage, two-dimensional conjugate structured codebook. Search efficiency is enhanced by a preselection process that constrains the exhaustive search to 32 of a possible 128 codevectors. The algorithm comes in two versions: the original G.729 (20 MIPS) and the less complex G.729 Annex A (11 MIPS). The algorithms are interoperable and the lower complexity algorithm has slightly lower quality. The MOS for the G.729 is 4.1 and for the G.729A 3.76. G.729 Annex B defines a silence compression algorithm allowing either the G.729 or the G.729A to operate at lower rates, thereby making them particularly useful in digital simultaneous voice and data (DSVD) applications. Also extensions to G.729 at 6.4 and 12 kbps are planned.

6.2.2. CELP Algorithms for PC Videoconferencing and Voice-over-IP Applications. The ITU G.723.1 [38] is a dual-rate speech coder algorithm intended for audio/videoconferencing/telephony over public phone (POTS) networks. G.723.1 is part of the ITU H.323 and H.324 audio/videoconferencing standards. The standard is dual rate 6.3 and 5.3 kbps. The excitation is selected using an analysis-by-synthesis process, and two excitation schemes are defined: the multipulse maximum-likelihood quantization (MP-MLQ) for the 6.3-kbps mode and the ACELP for 5.3 kbps. Ten short-term LP parameters are computed and vector-quantized as LSPs. A fifth-order LTP is used in this standard, and the LTP lag is determined using a closed-loop process searching around a previously obtained open-loop estimate. The LTP gains are vectorquantized. The high-rate MP-MLQ excitation involves matching of the LP residual with a set of impulses with restricted positions. The lower-rate ACELP excitation is similar but not identical to the excitation scheme used in G.729. G.723.1 provides a toll-quality MOS of 3.98 at 6.3 kbps and has a frame size of 30 ms with a lookahead of 7.5 ms. The estimated one-way delay is 37.5 mS. An option for variable-rate operation using a voice activity detector (silence compression) is also available. The Voice over IP (VoIP) forum, that is part of The International Multimedia Teleconferencing Consortium (IMTC), recommended G.723.1 to be the default audio codec for voice of the network (decision pending).

6.2.3. The ETSI GSM 6.60 Enhanced Full-Rate Standard. The enhanced full-rate (EFR) encoder was developed for use in the full-rate GSM standard [23]. The EFR is a 12.2-kbps algorithm with a frame of 20 ms and a 5-ms lookahead. A tenth-order short-term predictor is used, and its parameters are transformed to LSPs and encoded using split-vector quantization. The LTP lag is determined using a two-stage process where open-loop search provides an initial estimate which is then refined by closed-loop search around the neighborhood of the initial estimate. An algebraic codebook similar to that of the G.729 ACELP is also used. The algorithmic delay for the EFR is 25 ms and the MOS estimated to be around 4.1. The standard has provisions for a voice activity detector and an elaborate error protection scheme is also part of the standard. The North American PCS 1900 standard uses the GSM infrastructure and hence the EFR.

6.2.4. The Use of the EFR Algorithm in the IS641 TDMA Cellular/PCS Standard. The IS641 [25] is a 7.4-kbps EFR algorithm, also known as the Nokia/USH algorithm, and it is a variant of the GSM EFR. IS641 is the speech coding standard that is embedded in the IS136 Digital-AMPS (D-AMPS) North American digital cellular standard. EFR offers improved quality for this service relative to IS54. The algorithm is based on a 20-ms frame and 10th-order LP whose parameters are split-vector-quantized as LSPs. An algebraic codebook (ACELP) is used and the LTP lag is determined using open-loop search followed by closed-loop refinement with a fractional pitch resolution. The complexity of this coder is estimated at 14 MPS, and the Mean Opinion Score is estimated at 3.8.

6.2.5. The Relaxed CELP for the IS127 Enhanced Variable-Rate Coder (EVRC) for CDMA. The IS127 enhanced variable-rate coder (EVRC) [24] is based on relaxed CELP (RCELP), which uses interpolative coding methods [4] as a means for reducing further the bit rate and complexity in analysis-by-synthesis linear predictive coders. The EVRC encodes the RCELP parameters using a variable-rate approach. There are three possible bit rates for EVRC-8, 4, and 0.8 kbps-or after error protection, 9.6, 4.8, and 1.2 kbps, respectively. The rate is determined using a voice activity detection algorithm that is embedded in the standard. Rate changes can also be initiated on command from the network. At the lowest rate (0.8 kbps) the algorithm does not encode excitation information; hence the decoder essentially produces comfort noise. The other rates are achieved by changing the number of bits allotted to LP and excitation parameters. The algorithm uses a 20-ms frame, and each frame is divided in three 6.75ms subframes. Tenth-order short-term LP parameters are obtained using the Levinson-Durbin algorithm and splitvector encoded as LSPs. The fixed codebook structure is ACELP. The LTP parameters are estimated using generalized analysis-by-synthesis where instead of matching the input speech, a down-sampled version of a modified LP residual that conforms to a pitch contour is matched. The pitch contour is established using interpolative methods. The standard also specifies an FFT-based speech enhancement pre-processor that is intended to remove background noise from speech. The MOS for EVRC is 3.8 at 9.6 bps, and the algorithmic delay is estimated to be 25 ms.

6.2.6. The Japanese PDC Full-Rate and Half-Rate Standards. The Japanese Research and Development Center for Radio Systems (RCR) has adopted two algorithms for the personal digital cellular (PDC) full-rate (6.3 kbps) and the PDC half-rate (3.45 kbps) standards. The full-rate algorithm [33] is a variant of the IS54 VSELP algorithm described in a previous section. The half-rate PDC coder is a high-complexity pitch-synchronous innovation CELP (PSI-CELP) [39]. As the name implies, the codebooks of PSI-CELP depend on the pitch period. PSI-CELP defaults to periodic vectors if the pitch period is less than the frame size. The complexity of the algorithm is about 50 MIPS and the algorithmic delay is about 50 ms.

6.3. Third-Generation CELP for 3G Cellular Standards

The effort to establish wideband wireless cellular standards have driven further research and development toward algorithms that work at multiple rates and deliver significantly enhanced speech quality. These thirdgeneration algorithms are multimodal as they accommodate several different bit rates. This is consistent with the vision on wideband wireless standards [44] that will operate in different modes, with low mobility, high mobility, indoors, and so on. At least two algorithms have been developed and standardized for these applications. In Europe GSM is looking at the adaptive multirate coder [42,43], and in the United States the TIA has tested the *selectable mode vocoder* (SMV) [45,52] developed by Connexant. **6.3.1.** The Adaptive Multirate (AMR) Coder for GSM. An adaptive GSM multirate coder [42,43] has been adopted by ETSI for use in WCDMA. This is an ACELP algorithm that operates at multiple rates: 12.2, 10.2, 7.95, 6.7, 5.9, 5.15, and 4.75 kbps. The bit rate is adjusted according to the traffic conditions. The AMR is based on ACELP with 20-ms frame and 5-ms subframes. It uses a 10th-order short-term LPC and encodes LSPs using split vector quantization. At the highest bit rate it provides toll quality, and at the half rate it provides communications quality.

6.3.2. The Selectable Mode Vocoder. The SMV algorithm was developed to provide higher quality, flexibility, and capacity over the existing IS96C and IS127 EVRC CDMA algorithms. The SMV is based on 4 codecs: full rate at 8.5 kbps, half rate at 4 kbps, quarter rate at 2 kbps, and eighth rate at 800 bps. The full rate and half rate are based on the eXtended CELP (eX-CELP) algorithm, which is based on a combined closed-loop/openloop-analysis (COLA). In eX-CELP the signal frames are first classified as silence/background noise, nonstationary unvoiced, stationary unvoiced, onset, nonstationary voiced, and stationary voiced. The algorithm includes voice activity detection (VAD) followed by an elaborate frame classification scheme. Silence/background noise and stationary unvoiced frames are represented by spectrum modulated noise and coded at rate $\frac{1}{4}$ or $\frac{1}{8}$. The SMV uses four subframes for full rate and three subframes for half rate. The stochastic (fixed) codebook structure is also elaborate and uses subcodebooks, each tuned for a particular type of speech. The subcodebooks have different degrees of pulse sparseness (more sparse for noise like excitation). SMV scored as high as 4.1 MOS at full rate with clean speech.

7. SUMMARY

In this article, we provided an overview of some of the current linear predictive vocoder methods for speech and audio coding. Speech coding research has come a long way since the early 1990s, and several algorithms are rapidly finding their way in consumer products ranging from wireless cellular telephones to computer multimedia systems. Research and development in code excited linear prediction yielded several algorithms that have been adopted, or are strong candidates for adoption, in international wired and wireless communication standards from 16 down to 4.8 kbps. On the other hand, research activities are concentrating on the development of toll-quality algorithms that operate under 4 kbps. ITU has called for proposals for toll quality at 4 kbps for use in future videophones, personal communications, and third-generation cellular systems. Several proposals have been submitted [45-52] with none selected as of yet. Our coverage focussed on vocoder methods based on linear prediction because they have been the most popular in more recent low-rate communications and media standards. We must mention that in addition to the linear predictive coders, some frequency-domain methodologies have also been used in low-rate speech coding. In particular, sinusoidal analysis-synthesis schemes have been used as alternative to LPC. The sinusoidal model [8] represents speech by a linear combination of sinusoids. A low-rate sinusoidal transform coder (STC) has been considered in federal and TIA standardization competitions and is currently

considered for scalable wideband and high-fidelity applications. Another sinusoidal system, the *improved multiband excitation* (IMBE) coder [9,10], became part of the Australian (AUSSAT) mobile satellite standard and the International Mobile Satellite (INMARSAT-M) standard.

BIOGRAPHY

Andreas Spanias is professor in the Department of Electrical Engineering at Arizona State University (ASU). His research interests are in the areas of adaptive signal processing and speech/audio processing. While at ASU, he has developed and taught courses in DSP, adaptive signal processing, and speech coding. Andreas Spanias has received research contracts and grants from the National Science Foundation, Intel Corporation, Sandia National Labs, Motorola Inc., and Active Noise and vibration Technologies. He has also consulted with Inter-Tel Communications, Texas Instruments, and the Cyprus Institue of Neurology and Genetics. He authored the J-DSP educational software package (ISBN 0-9724984-0-0) that is used for on-line labs in DSP. He is a Senior Member of the IEEE and has served as a member on two IEEE technical comittees of the IEEE Signal Processing Society (SPS). He has also served as associate editor of the IEEE Transactions on Signal Processing and the IEEE Signal Processing Letters. He co-chaired the 1999 International Conference on Acoustics Speech and Signal Processing (ICASSP-99) in Phoenix, and he currently is the IEEE SPS Vice-President for Conferences. Andreas Spanias is the co-recipient of the 2002 IEEE Donald G. Fink prize paper award for the IEEE Proceedings manuscript entitled "Perceptual Coding of Digital Audio."

BIBLIOGRAPHY

- 1. M. R. Schroeder and B. Atal, Code-excited linear prediction (CELP): High quality speech at very low bit rates, *Proc. ICASSP'85*, Tampa, April 1985, p. 937.
- 2. S. Singhal and B. Atal, Improving the performance of multipulse coders at low bit rates, *Proc. ICASSP'84*, 1984, p. 1.3.1.
- B. Atal and J. Remde, A new model for LPC excitation for producing natural sounding speech at low bit rates, *Proc. ICASSP'82*, April 1982, pp. 614–617.
- W. B. Kleijn et al., Generalized analysis-by-synthesis coding and its application to pitch prediction, *Proc. ICASSP'92*, 1992, pp. 1337–1340.
- 5. P. Kroon, E. Deprettere, and R. J. Sluyeter, Regular-pulse excitation—a novel approach to effective and efficient multipulse coding of speech, *IEEE Trans.* ASSP-34(5) (Oct. 1986).
- I. Gerson and M. Jasiuk, Vector sum excited linear prediction (VSELP) speech coding at 8 kbits/s, *Proc. ICASSP'90*, New Mexico, April 1990, pp. 461–464.
- 7. A. McCree and T. Barnwell III, A new mixed excitation LPC vocoder, *Proc. ICASSP'91*, Toronto, May 1991, pp. 593–596.
- R. McAulay and T. Quatieri, Low-rate speech coding based on the sinusoidal model, in S. Furui and M. M. Sondhi, eds., *Advances in Speech Signal Processing*, Marcel Dekker, New York, 1992, Chap. 6, pp. 165–207.
- D. Griffin and J. Lim, Multiband excitation vocoder, *IEEE Trans.* ASSP-36(8): 1223 (Aug. 1988).

- J. Hardwick and J. Lim, The application of the IMBE speech coder to mobile communications, *Proc. ICASSP'91*, May 1991, pp. 249–252.
- P. Noll, Digital audio coding for visual communications, Proc. IEEE 83(6): 925–943 (June 1995).
- T. Painter and A. Spanias, Perceptual coding of digital audio, *Proc. IEEE* 88(4): 451–513 (April 2000).
- ITU Study Group 15 Draft Recommendation G.729, Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), International Telecommunication Union, 1995.
- TIA/EIA-PN 2398 (IS54), The 8 kbit/s VSELP Algorithm, 1989.
- GSM 06.10, GSM Full-Rate Transcoding, Technical Report, Version 3.2, ETSI/GSM, July 1989.
- Federal Standard 1015, Telecommunications: Analog to Digital Conversion of Radio Voice by 2400 Bit/Second Linear Predictive Coding, National Communication System—Office Technology and Standards, Nov. 1984.
- J. Makhoul, Linear prediction: A tutorial review, Proc. IEEE 63(4): 561-580 (April 1975).
- J. Markel and A. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.
- N. Benevuto et al., The 32Kb/s coding standard, AT&T Tech. J. 65(5): 12–22 (Sept.–Oct. 1986).
- ITU Recommendation G.726 (formerly G.721), 24, 32, 40 kb/s Adaptive Differential Pulse Code Modulation (ADPCM), Blue Book, Vol. III, Fascicle III.3, Oct. 1988.
- A. Spanias, Speech coding: A tutorial review, *Proc. IEEE* 82(10): 1541-1582 (Oct. 1994).
- I. Boyd and C. Southcott, A speech codec for the skyphone service, Br. Telecommun. Techn. J. 6(2): 51–55 (April 1988).
- GSM 06.60, GSM Digital Cellular Communication Standards: Enhanced Full-Rate Transcoding, ETSI/GSM, 1996.
- TIA/EIA/IS-127, Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, TIA, 1997.
- TIA/EIA/IS-641, Cellular/PCS Radio Interface—Enhanced Full-Rate Speech Codec, TIA, 1996.
- P. Kroon and B. Atal, Pitch predictors with high temporal resolution, *Proc. ICASSP'90*, New Mexico, April 1990, pp. 661–664.
- 27. INMARSAT-B SDM Module 1, Appendix I, Attachment 1, MAC/SDM/BMOD1/ATTACH/ISSUE 3.0.
- J. Campbell and T. E. Tremain, Voiced/unvoiced classification of speech with applications of the U.S. government LPC-10e algorithm, *Proc. ICASSP'86*, Tokyo, 1986, pp. 473–476.
- T. E. Tremain, The government standard linear predictive coding algorithm: LPC-10, Speech Technol. 40–49 (April 1982).
- J. Makhoul et al., A mixed-source model for speech compression and synthesis, J. Acous. Soc. Am. 64: 1577–1581 (Dec. 1978).
- J. Campbell, T. E. Tremain, and V. Welch, The proposed federal standard 1016 4800 bps voice coder: CELP, Speech Technol. 58-64 (April 1990).
- 32. Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4800 Bit/Second Code

Excited Linear Prediction (CELP), National Communication System—Office Technology and Standards, Feb. 1991.

- 33. I. Gerson, Vector sum excited linear prediction (VSELP) speech coding for Japan digital cellular, paper presented at meeting of IEICE, RCS90-26, Nov. 1990.
- GSM 06.20, GSM Digital Cellular Communication Standards: Half Rate Speech; Half Rate Speech Transcoding, ETSI/GSM, 1996.
- 35. ITU Draft Recommendation G.728, Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction (LD-CELP), 1992.
- J. Chen et al., A low-delay CELP coder for the CCITT 16 kB/s speech coding standard, *IEEE Trans. Selec. Areas Commun.* (Special Issue on Speech and Image Coding; N. Hubing, ed.) 830-849 (June 1992).
- TIA/EIA/IS-96, QCELP, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, TIA 1992.
- ITU Recommendation G.723.1, Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s, Draft 1995.
- T. Ohya, H. Suda, and T. Miki, The 5.6 kb/s PSI-CELP of the half-rate PDC speech coding standard, *Proc. IEEE Vehicular Technology Conf.*, 1993, pp. 1680–1684.
- 40. ITU Recommendation G.722, 7 KHz Audio Coding within 64 kbits/s, Blue Book, Vol. III, Fascicle III, Oct. 1988.
- INMARSAT Satellite Communications Services, INMAR-SAT-M System Definition, Issue 3.0—Module 1: System Description, Nov. 1991.
- 42. ETSI AMR Qualification Phase Documentation, 1998.
- R. Ekudden, R. Hagen, I. Johansson, and J. Svedburg, The adaptive multi-rate speech coder, *Proc. IEEE Workshop on* Speech Coding, 1999, pp. 117–119.
- D. N. Knisely, S. Kumar, S. Laha, and S. Navda, Evolution of wireless data services: IS-95 to cdma2000, *IEEE Commun. Mag.* 140–146 (Oct. 1998).
- Conexant Systems, Conexant's ITU-T 4 kbit/s Deliverables, ITU-T Q21/SG16 Rapportuer meeting, AC-99-20, Sept. 1999.
- Matsushita Electric Industrial Co. Ltd., High Level Description of Matsushita's 4-kbit/s Speech Coder, ITU-T Q21/SG16 Rapportuer meeting, AC-99-19, Sept. 1999.
- Mitsubhishi Electric Corp., High Level Description of Mitsubhishi 4 kb/s Speech Coder, ITU-T Q21/SG16 Rapportuer meeting, AC-99-016, Sept. 1999.
- NTT, High Level Description of NTT 4 kb/s Speech Coder, ITU-T Q21/SG16 Rapportuer meeting, AC-99-17, Sept. 1999.
- Rapporteur (Mr. Paul Barrett, BT/UK), Q.21/16 Meeting Report, ITU-T Q21/SG16 Rapportuer meeting, Temporary Document E (3/16), Geneva, Feb. 7–18, 2000.
- Texas Instruments, High Level Description of TI's 4 kb/s Coder, ITU-T Q21/SG16 Rapportuer meeting, AC-99-25, Sept. 1999.
- Toshiba Corp., Toshiba Codec Description and Deliverables (4 kbit/s Codec), ITU-T Q21/SG16 Rapportuer meeting, AC-99-15, Sept. 1999.
- 52. Y. Gao et al., The SMV algorithm selected for TIA and 3GPP2 for CDMA applications, *ICASSP*'2001, Salt Lake City, May 5–12, 2002, Vol. 2.

WAVEFORM CODING

GÜNEŞ KARABULUT ABBAS YONGAÇOĞLU University of Ottawa School of Information Technology and Engineering Ottawa, Ontario, Canada

1. INTRODUCTION

A signal is an entity that changes with time and contains some information. If a signal is continuous in both time and amplitude, it is referred to as an *analog signal* (*waveform*) and can be represented by a physical quantity (e.g., voltage, current, pressure) proportional to it. If it is discrete both in time and amplitude, then it is referred to as a *digital signal* and can be represented in bits. Conversion from one form to the other is performed whenever necessary. For example human speech is an analog signal but what we hear from compact disks are digital signals since they are stored and processed as ones and zeros.

Digital representation of analog signals offers many advantages. Digital signals are less sensitive to various transmission impairments and noise, easier to store and regenerate, and suitable for encryption for greater security. It is easy to multiplex various forms of digital information. Digital signals also enable unification of transmission and switching functions in communications and permit error protection.

Waveform coding is the process of describing analog signals in a digital form. It can also be viewed as the process of associating a mathematical representation with a waveform. The major goal of waveform coding is to obtain the minimum data rate for a given amount of distortion; or conversely, to represent a waveform at a given data rate while causing the minimum amount of distortion.

The area of waveform coding has been very active since the early 1940s, and it is still going strong with the spread of digitization. In this article we will first discuss the basic principles of sampling and quantization, and will then present temporal and spectral waveform coding techniques.

2. DIGITIZATION

A source generates the information to be transmitted. If a source is producing waveforms, then it is an analog source, and the source encoder converts those waveforms to a digital form suitable for transmission or storage. A waveform encoder follows an analog source, and its location in a generic digital communication system is shown in Fig. 1. If the source is already producing digital symbols, then the source encoder compresses the information by taking advantage of its statistical properties.

The source decoder tries to reconstruct the input waveform from the received digital data. The semblance of the input signal to its received estimate obtained from the source decoder determines the performance of the waveform coding system. This performance depends mainly on two factors: quantization noise and channel impairments. *Quantization noise* is introduced in the waveform encoder, and it is the result of representing infinite number of amplitudes an input signal can take by using only a finite number of amplitudes. The degradation due to quantization noise ratio (SQNR), as will be defined in later sections. SQNR depends on the input waveform properties and the specific source encoding technique employed.

Channel impairments are experienced while modulated waveform is transmitted through a noisy communication channel. A topic that has gained popularity is *channel optimized coding*, which factors in the noisy nature of the channel in the source coding process [1].

Waveforms are continuous in amplitude and time. In order to produce their digital equivalents, we need to make them discrete in both amplitude and time. The process of time discretization (for two-dimensional signals this corresponds to space discretization) is called *sampling*, and the process of amplitude discretization is called quantization. The circuit that performs sampling and quantization is often referred to as an analog-to-digital converter (ADC). Block diagram of an ADC is shown in Fig. 2. The prefilter shown is added to bandlimit input waveforms as will be explained later. A waveform x(t), and the corresponding prefilter, sampler, and quantizer outputs are shown in Fig. 3. When sampled at or above the Nyquist rate (defined in the next section), the original analog signal can always be fully recovered from its samples.

2.1. Sampling and Reconstruction

2.1.1. Sampling. Sampling is the process of converting a waveform to a series of samples in time. A sample



Figure 1. A digital communication system.



Figure 2. Analog-to-digital converter (ADC).



Figure 3. Signal samples of ADC: (a) input signal, x(t) amplitude-time plot; (b) prefilter output that band-limits the input signal; (c) sampler output $x_{\delta}(t)$; (d) quantizer output, x.

is a measure of amplitude of a waveform evaluated over a period of time. When this period is constant for each sample (i.e., samples are equally spaced in time), the process is called *uniform sampling*. The period is termed the sampling period or sampling interval, T_s . The reciprocal of T_s is called the sampling rate or sampling frequency and denoted by f_s .

In the following analysis uniform sampling will be considered because of its widespread usage and convenience. To explain the sampling process, we must first introduce the concept of a band-limited waveform.

2.1.2. Band-Limited Waveform. Let x(t) be a real valued waveform with finite energy

$$\int_{-\infty}^{+\infty} |x(t)|^2 dt < \infty, \tag{1}$$

so that x(t) is Fourier transformable. Denoting the Fourier transform of x(t) as X(f), the Fourier transform pair

equations are given by

$$X(f) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi f t} dt \Leftrightarrow x(t) = \int_{-\infty}^{+\infty} X(f)e^{j2\pi f t} df \quad (2)$$

A waveform band-limited to *W* hertz or $\Omega_W = 2\pi W$ radians per second (rad/s) is defined as

$$X(f) = 0; \quad |f| \ge W = \frac{\Omega_W}{2\pi} \tag{3}$$

2.1.3. Sampling Theorem. Let the finite energy signal x(t) be also band-limited, and let $x_{\delta}(t)$ denote the sampled version of x(t). Defining the Dirac delta function $\delta(t)$ as $\delta(t) = 0$, for $t \neq 0$ and $\int_{-\infty}^{+\infty} \delta(t)dt = 1$, the relationship between x(t) and $x_{\delta}(t)$ is given by

$$x_{\delta}(t) = \sum_{k=-\infty}^{+\infty} x(t)\delta(t - kT_s)$$
(4)

where $\delta(t - kT_s)$ is the Dirac delta function positioned at $t = kT_s$. Through the sifting property of Dirac delta function we can modify Eq. (4) as

$$x_{\delta}(t) = \sum_{k=-\infty}^{+\infty} x(kT_s)\delta(t - kT_s)$$
(5)

This equation represents a modified impulse train with weights of $x(kT_s)$ at $t = kT_s$. Therefore uniform sampling can be visualized as the multiplication of x(t) by an infinitely long train of impulse functions that are T_s seconds apart.

The frequency domain view of the sampled signal becomes

$$X_{\delta}(f) = \sum_{k=-\infty}^{+\infty} x(kT_s) e^{-j2\pi k/T_s}$$
(6)

where we can see that the uniform sampling process results in a periodic spectrum with a period of $1/T_s$. If we choose the sampling rate as $T_s = 1/2W$, which is named as the Nyquist rate, we obtain the spectrum as

$$X_{\delta}(f) = \sum_{k=-\infty}^{+\infty} x\left(\frac{k}{2W}\right) e^{-j\pi k f/W}$$
(7)

If we choose a sampling period larger than 1/2W (i.e., $f_s < 2W$), we would observe the effect of aliasing. In this case the sampler output is referred to as *undersampled*, and the original waveform x(t) cannot be recovered from $x_{\delta}(t)$. If we choose a sampling period smaller than 1/2W (i.e., $f_s > 2W$), the original waveform x(t) can be



Figure 4. Sampling process in frequency domain: (a) frequency spectrum of x(t), X(f); (b) sampled at Nyquist rate; (c) undersampled, aliasing present; (d) oversampled.

recovered, but we will observe bandwidth expansion due to oversampling.

In Fig. 4, the effects of the sampling period in the frequency domain are shown. Frequency spectrum of x(t), X(f) is shown in Fig. 4a. The spectrum shown in Fig. 4b is sampled at the Nyquist rate. In Fig. 4c the signal is undersampled and we can see the effect of aliasing, and in Fig. 4d the signal is oversampled.

The fourier transform of x(t) can also be expressed as

$$X_{\delta}(f) = f_s \sum_{k=-\infty}^{+\infty} X(f - kf_s) = f_s X(f) + \sum_{\substack{k=-\infty\\k\neq 0}}^{+\infty} X(f - kf_s)$$
(8)

Hence

$$X_{\delta}(f) = f_s X(f), \quad -W < f < W \tag{9}$$

The reconstruction method of x(t) from this equation will be explained in the following section. **2.1.4.** Reconstruction. To recover x(t) from $x_{\delta}(t)$, we need to use a reconstruction filter, which has the following properties:

$$H_R(f) = \frac{1}{f_s} = T_s \quad |f| \le W$$

$$H_R(f) = 0 \qquad |f| > W$$
(10)

The corresponding impulse response of the reconstruction filter is

$$h_R(t) = \frac{\sin(\pi t/T_s)}{\pi t/T_s} = \operatorname{sinc}\left(\frac{\pi t}{T_s}\right)$$
(11)

The multiplication of $X_{\delta}(f)$ by $H_R(f)$, is equivalent to convolution of $x_{\delta}(t)$ with $h_R(t)$ in the time domain. Thus, we obtain the interpolation formula as

$$x_{\delta}(t)^* h_R(t) = \sum_{k=-\infty}^{+\infty} x(kT_s) \operatorname{sinc}\left(\frac{t}{T_s} - k\right) = x(t)$$
(12)

where * denotes the convolution operation.

When x(t) is band-limited and the sampling rate exceeds 2W, the sampling and reconstruction operations are error-free. In practice, in order to recover the input waveform exactly from its samples, the waveform must be band-limited. To render x(t) band-limited, a prefilter (also known as the antialiasing filter) can be inserted before the sampler as shown in Fig. 2. The aim of this filter is to prevent or minimize the effects of aliasing. The distortion caused by the prefilter with an appropriately chosen bandwidth is less objectionable than the effect of aliasing. For band-limited signals, prefiltering operation does not have a distorting effect. The input signal x(t)and its prefiltered version are shown in Figs. 3a and 3b, respectively. The sampled version of prefiltered signal, $x_{\delta}(t)$, is shown in Fig. 3c. The quantized version of $x_{\delta}(t)$, **x**, is shown in Fig. 3d.

Another precaution to prevent aliasing and to fully recover the original signal is sampling at a frequency slightly higher than the Nyquist frequency of 2*W*.

2.2. Quantization

The continuous amplitude of a sample can take on a value from an infinitely large set. However, to represent this value digitally, a finite set of discrete amplitude values are used. This process is called *quantization*. The simplest form of quantization is rounding off a number. Unlike sampling, the quantization is an irreversible process; that is, from a quantized value we cannot go back to the original value. Hence it can be viewed as a type of lossy data compression.

Two major parameters in the quantization process are the data rate and distortion. The unit of data rate can be bits per second (bps) or bits per sample. To save transmission and/or storage resources, quantization at the minimal rate is highly desirable. However, rate and distortion are limited by Shannon's rate distortion theory [2]. Hence the other view of the rate distortion tradeoff is to achieve the minimum distortion for a given rate. Distortion can be defined as a subjective
quantity, involving the ratings that are given by experts such as mean opinion score, or it can be an objective quantity defined in mathematical terms. An objective distortion definition resulting from the quantization of signal amplitudes is

$$D = \int_{-\infty}^{+\infty} F[x_{\delta}(t) - \mathbf{x}] \, dx \tag{13}$$

where $F[\cdot]$, denotes the desired error function.

A quantizer can be optimized according to the probability density function (pdf) of the input; therefore it is waveform specific.

The difference between the input and the output of a quantization process is defined as the *quantization error*, e(t), and is given by

$$e(t) = x_{\delta}(t) - \mathbf{x} \tag{14}$$

Performance of a quantizer is often measured by the *signal-to-quantization noise ratio* (SQNR), which is defined as

$$SQNR = \frac{P_x}{\sigma_e^2}$$
(15)

where P_x denotes the input signal power and σ_e^2 denotes the variance of the quantization error, e(t).

Two major quantization techniques are scalar quantization and vector quantization.

2.2.1. Scalar Quantization (SQ). Scalar quantization (SQ) is the mapping of a sample to the nearest quantization level. *Quantization interval* is the range of amplitudes that are mapped to the same quantization level. The difference between two quantization levels is referred to as the step size. These concepts are shown in Fig. 5. A quantization level should be chosen for each quantization interval. In an *n*-bit scalar quantizer each sample is represented by *n* bits, and there can be 2^n different quantization levels.

2.2.1.1. Uniform Quantization. In uniform quantization, the quantization intervals are of equal length, and each quantization level is chosen as the midpoint of a quantization interval. The distortion resulting from the



Figure 5. A uniform eight-level midrise quantizer.

quantization process is directly proportional to the square of the step size and therefore is inversely proportional to the number of levels, n. A uniform eight-level midrise quantizer is shown in Fig. 5.

2.2.1.2. Adaptive Quantization. Adaptive quantization is used for samples with dynamically changing range of amplitudes. They have uniform step sizes, but the length of step sizes changes according to the dynamic range of the input waveform. There are two major approaches [3]:

- 1. Offline Adaptive (Forward-Adaptive) Approach. In this type of quantization, source output is first divided into blocks. Each input block is individually analyzed and distinct quantizer parameters are assigned. Quantizer parameters should be transmitted as side information.
- 2. Online Adaptive (Backward-Adaptive) Approach. In this method, adaptation is based on the quantizer output. Since parameters are available both at the transmitter and receiver, no feedback link is necessary.

2.2.1.3. Nonuniform Quantization. Uniform quantization is optimum when the input signal levels are uniformly distributed. Often the distribution of the input signal is nonuniform, and using a nonuniform quantizer instead of a uniform quantizer results in a considerable performance improvement. In nonuniform quantization, the quantizer step sizes are not equal. Usually step sizes close to the mean signal level are smaller and the step sizes become larger as the input signal deviates further from the mean. Having the same number of quantization levels, nonuniform quantizers result in a lower average distortion than uniform quantizers. The expense associated with nonuniform quantizers is their more complex structure.

There are various kinds of nonuniform quantization schemes. One method is to optimize the quantizer levels to produce minimum quantization error according to the source statistics. This can be achieved by compressing the input amplitudes according to a PDF, and then applying the resulting signal to a uniform quantizer. At the decoder a corresponding expander is utilized in the reconstruction process. The mappings that are performed in the blocks are also shown in Fig. 6. This type of quantizer is termed compander, because of *compressor* and expander that exist in the system. A major application area of companders is commercial telephony. There are two frequently used companding techniques named μ law and A law, which are used in the United States and Europe, respectively [4].

2.2.2. Vector Quantization (VQ). Although Shannon has stated that given an average distortion D and rate R(D), a waveform can be reconstructed with a distortion arbitrarily close to D, this is not practically achievable with scalar quantization. Vector quantization (VQ) takes advantage of processing many samples at once for a more accurate representation of source output. VQ works best when input blocks are correlated, but even when the quantized samples are independent VQ performs better than scalar quantization [4].



Figure 6. Compander.

In VQ, instead of processing one sample at a time, a group of samples (called a *vector*) is mapped to a codebook index. A *codebook* is a set consisting of a finite number of vectors formed in such a way that each input vector has a corresponding output.

The dimension of a vector quantizer is defined as the number of samples in the vector. The rate of the vector quantizer is defined as

$$R = \frac{\log_2 n}{L} \text{ bits per sample}$$
(16)

where n is the size of VQ codebook and L is the dimension of the quantizer [3].

VQ structure introduces increased implementation complexity. To alleviate this problem, tree-structured VQ can be used. Details of various aspects of vector quantization can be found in the literature [4-6].

3. CODING TECHNIQUES

Waveform coding techniques can be classified into several categories. In this article we discuss only major temporal and spectral coding techniques. Further details on various coding techniques can be found in the literature [4,5,7]. Another important category is model based coding, which is not covered here.

3.1. Temporal Coding Techniques

In temporal coding techniques the waveform coding process is performed in the time domain. In the following sections we will discuss the basic temporal coding techniques: pulse code modulation (PCM), differential PCM (DPCM), and delta modulation (DM). Details on these and other temporal coding methods can be found in the literature [4,5,7].

3.1.1. Pulse Code Modulation (PCM). Pulse code modulation (PCM) has gained popularity with digital telephony. It is the most straightforward method of digitization as we saw in the last section. Conceptually the encoder performs three functions: (1) sampling the input waveform, (2) quantizing the samples, and (3) representing each quantizer level with a binary index.

A typical PCM system is shown in Fig. 7. An input signal is first prefiltered to ensure band limitation, and then is sampled and quantized. Next, the quantized amplitude values are mapped to electrical signals suitable for transmission over a particular channel. The received signals are demodulated and passed through the reconstruction filter. Finally, an estimate of the input waveform is received by the destination.

Assuming that the quantization noise is uniformly distributed over the quantization intervals, it has been



Figure 7. A typical PCM system.

shown [3] that the performance with an n-bit uniform quantizer can be evaluated by

$$(SQNR)_{dB} = 6.02n + \alpha \tag{17}$$

where $\alpha = 4.07$ for peak SQNR and $\alpha = 0$ for average SQNR. Hence each additional bit improves the performance by about 6 dB.

PCM has several advantages, including immunity to channel impairments such as noise and interference and ease of implementation. Because of its widespread use, PCM has become a standard for comparison of data compression schemes. PCM systems using companding generally provide toll quality speech at a rate of 64 Kbps.

3.1.2. Differential Pulse Code Modulation (DPCM). PCM is designed without taking into account any correlation between successive samples. In practice the successive samples of most waveforms are highly correlated. The key idea behind differential pulse code modulation (DPCM) is to exploit this correlation.

A typical DPCM system is shown in Fig. 8. In this system, an estimate of the current sample value is produced by the predictor through the feedback loop that is located in the encoder part. Next, this estimate is subtracted from the current sample value; the difference is quantized, modulated, and transmitted through the channel. Then in the receiver, demodulated symbols are recovered through the second feedback loop located at the receiver. In this loop, the predicted value is added to the quantized prediction error. The variance of the difference between a sample and its estimate is smaller than or equal to the variance of the original signal. Therefore the quantization noise power of a DPCM system is smaller than that of a PCM system. This results in a higher SQNR than that of a PCM system for the same data rate or the same SQNR value can be achieved with a lower data rate. A typical DPCM system provides toll-quality speech at a rate of 32-48 kbps.

Other versions of the DPCM system where predictor parameters change according to the statistics of the input signal and other parameters can provide additional improvements [5,7].

3.1.3. Delta Modulation (DM). Delta modulation (DM) is a simplified version of DPCM. DM is also based on the observation that input signal amplitudes generally do not make very sudden changes. The key point in DM is oversampling the input waveform with a sampling frequency much higher than the Nyquist rate, and increasing the correlation between the samples. DM can be visualized as a derivative approximation of an input waveform.

DM has almost the same structure as DPCM system shown in Fig. 8. The only change required for DM is to replace the predictor by a unit delay. In DM, following the input waveform within a $\pm \Delta$ range, a quantized signal is produced, as shown in Fig. 9. This type of approximation is known as "staircase approximation." When the ratio of step size to sampling period (i.e., Δ/T_s) is larger than the maximum slope of the input waveform, then slope overload occurs. The resulting distortion is named as slope overload distortion. The second type of quantization noise appears in DM when Δ is larger than the slope of the waveform. This is termed the granular noise. Both noise types are depicted in Fig. 9. The most crucial parameter in DM is Δ , also known as the *step size*. Large values of Δ permit the approximation of rapid changes, whereas smaller values of Δ help reduce the effect of granular noise. DM systems generally provide a rate of 32-64 kbps for toll-quality speech.

To reduce the effects of the abovementioned quantization noise types, adaptive delta modulation (ADM) can be used. In ADM Δ varies according to the input waveform. Details on ADM can be found in the treatise by Jayant [5].

3.2. Spectral Coding Techniques

In temporal coding techniques, the input is a full-band waveform, that is, not filtered into smaller frequency bands. In spectral waveform coding techniques, the input signal is divided into several frequency components and



Figure 8. A typical DPCM system.



Figure 9. Noise in a DM system.

each component is coded separately. Decomposing the source output into different frequency bands (subbands) is performed by digital filters. Each filter output component is quantized separately so that the resulting quantization noise is contained within its band.

Spectral coding techniques take advantage of human perception of signals such as image or speech. Human senses cannot detect quantization distortion at all frequencies with the same precision. Spectral coding techniques utilize dynamic bit allocation by assigning more bits to more important frequency components, and less bits to the others. Spectral coding techniques also increase efficiency by removing the redundancy between input samples, thus providing uncorrelated channel inputs.

Spectral coding techniques can be divided into two major classes: subband coding (SBC) and transform coding (TC). In practice SBC is generally used for audio signals [5], and TC is dominantly used for image signals [8].

3.2.1. Subband Coding (SBC). Subband coding (SBC) is a waveform coding method that first decomposes an input waveform into different subbands and then operates on them. As shown in Fig. 10, division of the signal into subbands is performed with the help of an analysis filterbank. Filter outputs are generally translated to low-pass signals by an operation equivalent to single-sideband

modulation. Then the outputs are sampled at Nyquist rate and encoded in the time domain, with the help of one of the temporal coding techniques such as PCM or DPCM. The coding technique for each branch is chosen according to the properties of the corresponding band, or a perceptual criterion. The resulting signals are multiplexed and transmitted through the channel. At the receiver part, the received signal is demultiplexed and decoded. The resulting signal is passed through a synthesis filterbank and the outputs are summed to form the reconstructed signal.

Quadrature mirror filters (QMFs) are the popular choice for the filterbanks. These filters obey certain symmetry conditions and can achieve perfect alias cancellation.

Depending on the nature of a particular application, SBC can operate over equal-width subbands or unequalwidth subbands. In speech coding applications usually four to eight subbands are used. A 32-kbps subband coder can provide toll-quality speech.

In SBC, since each filter output is encoded and quantized separately, the system can control and distribute the quantization noise across the spectrum. Therefore the quantization noise in each band can be adjusted independently from that in other bands. Different number of bits can be assigned to each subband. Also errors due to channel impairment affect only the outputs of corresponding frequency bands. Combination of subband coding with adaptive prediction systems can improve system performance [5].

3.2.2. Transform Coding (TC). Transform coding (TC) is also known as *block transform coding* or *block quantization*. As shown in Fig. 11, a transform coder encodes a block of a discrete input sequence according to a bit allocation scheme, derived from the perceptual significance of components. The input to the linear transformer can be obtained with a high-resolution temporal coder.

The efficiency of a TC technique depends on the type of linear transform applied and the bit allocation scheme used in the quantization of transform coefficients.



Figure 10. A typical subband coder.



Figure 11. A typical transform coder.

The main objective of the linear transformation in TC is to provide uncorrelated samples. In this sense, the *Karhunen-Loeve transform* (KLT), also known as the *Hotelling transform* or the *eigenvalue transform*, is optimum since it transforms discrete variables into uncorrelated coefficients. KLT is based on the statistical properties of vector representations of the input signal and minimizes the mean-square error between the input and its approximation. KLT has the maximum capacity to perform data compression.

A major disadvantage of the KLT is the high computational complexity. This problem can be solved by using a suboptimal but more practical transform method such as discrete Fourier transform (DFT), discrete cosine transform (DCT), discrete Hadamard transform (DHT), or discrete Walsh transform (DWT).

An important design issue in TC is how to implement the bit allocation scheme. Optimum bit allocation, which yields the best SQNR, depends on the variance of quantization coefficients but it is too complex for most practical purposes.

Zonal sampling is a simple bit allocation method. In zonal sampling zero bits are allocated to less important (i.e., out-of the zone) coefficients. In order to obtain a good performance from zonal sampling, the input power spectral density should be highly structured.

A more complex version of zonal sampling is threshold sampling, where a zone is defined as a variable region depending on the amplitudes of the coefficients.

TC provides a better performance than does SBC at the expense of increased coding delay and more complex structure. The number of subbands in TC is typically greater than the number of subbands used in SBC. This number is also referred to as the order of the transform.

A more complex version of TC is adaptive transform coding (ATC). It includes adaptive bit allocation from window to window while keeping the total data rate constant. ATC results in a significant increase in the SQNR with the expense of further increased complexity. More details on ATC can be found in Refs. 5 and 7.

BIOGRAPHIES

Güneş Karabulut was born in Istanbul, Turkey in 1979. She received the B.Sc. degree in Electrical and Electronics Engineering from Boğaziçi University, Istanbul, Turkey and the M.A.Sc. degree in Electrical Engineering from the University of Ottawa, Ontario, Canada. Currently she is pursuing the Ph.D. degree at the University of Ottawa.

From 1999 to 2000 she worked on motion estimation algorithms in Boğaziçi University Signal and Image Processing Laboratory. Since September 2000 she is employed as a Research Assistant at Communications and Signal Processing Group, University of Ottawa. Her research interests include coding theory, coded modulation schemes, and image coding. Ms. Karabulut is a member of IEEE Information Theory Society. **Abbas Yongaçoğlu** received the B.Sc. degree from Boğaziçi University, Turkey, in 1973, the M.Eng. degree from the University of Toronto, Canada, in 1975, and the Ph.D. degree from the University of Ottawa, Canada, in 1987, all in Electrical Engineering.

He worked as a researcher and a system engineer at TUBITAK Marmara Research Institute in Turkey, Philips Research Labs in Holland and Miller Communications Systems in Ottawa. In 1987 he joined the University of Ottawa as an assistant professor. He became an associate professor in 1992, and a full professor in 1996. His area of research is digital communications with emphasis on modulation, coding, equalization and multiple access for wireless and high speed wireline communications.

BIBLIOGRAPHY

- N. Farvardin and V. Vaishampayan, Optimal quantizer design for noisy channels: An approach to combined source channel coding, *IEEE Trans. Inform. Theory* **33**(6): 827–838 (Nov. 1987).
- C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE Conv. Rec.* 7: 142–163 (1959).
- T. S. Rappaport, Wireless Communication: Principles & Practice, Prentice-Hall, Englewood Cliffs, NJ, 1996.
- 4. K. Sayood, Introduction to Data Compression, Morgan Kaufmann, San Francisco, 2000.
- N. S. Jayant, Digital Coding of Waveforms Principles and Applications to Speech and Video, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- R. M. Gray, Vector quantization, *IEEE Acoust. Speech Signal Process. Mag.* 1(2): 4–29 (April 1984).
- J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 2001.
- R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison Wesley, New York, 1993.

FURTHER READING

- Gibson J. D., ed., The Mobile Communication Handbook, CRC Press, Boca Raton, FL, 1996.
- Haykin S., Communication Systems, Wiley, New York, 1994.
- Ortega A. and K. Ramchandran, Rate-distortion methods for image and video compression, *IEEE Signal Process. Mag.* 15(6): 23-50 (Nov. 1998) (gives details on rate distortion theory).
- Effros M., Optimal modeling for complex system design, *IEEE Signal Process. Mag.* **15**(6): 51–73 (Nov. 1998) (gives details on rate distortion theory).
- Goyal V. K., Theoretical foundations of transform coding, IEEE Signal Process. Mag. 18(5): 9–21 (Sept. 2001).
- Oppenheim A. V., R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1999 (an excellent source for more details on sampling and reconstruction processes).

WAVELENGTH-DIVISION MULTIPLEXING OPTICAL NETWORKS

EYTAN MODIANO Massachusetts Institute of Technology Cambridge, Massachusetts

1. INTRODUCTION

Communication networks were first developed for providing voice telephone service. Early networks were deployed using copper wire as the medium over which traffic was sent in the form of electromagnetic waves. As demand for communication increased, networks began to use optical fiber cables over which information was sent in the form of lightwaves. Thanks to the relatively low attenuation losses of optical fiber, transmitting information over fiber allowed for a significant increase in the transmission capacity of networks. For example, while transmission over copper was limited to only a few tens of thousands of bits per second (kbps), optical fiber transmission enabled data rates exceeding hundreds of millions of bits per second (Mbps). However, the relatively recent development of the Internet has resulted in a tremendous increase in demand for transmission capacity. More recently, developments in optical transmission technology have achieved data rates that exceed many billions of bits per second (Gbps). Even with these enormous data rates, demand still far exceeds the available network capacity. As a result, telecommunication equipment companies are constantly trying to develop new technology that can increase network capacity at reduced costs.

While fiberoptic technology resulted in a significant increase in a network's "bandwidth," or the amount of information that the network could send, the creation of the Internet resulted in an even greater demand for bandwidth. As demand for network capacity increased, service providers exhausted their available transmission capacity. One approach to alleviating fiber exhaust is to deploy additional fiber. This solution, however, is not always economically feasible. As a result, new technologies were developed to increase the transmission capacity of existing fiber.

The simplest approach is to increase the rate of transmission over the fiber (i.e., sending more bits per second). Since 1980, fiber transmission rates have increased from a few Mbps to nearly 100 Gbps. Since most users rarely need such high data rates, a network technology called *synchronous optical networks* (SONET)

was developed to allow users to share the capacity of a fiber [1].

SONET is a technology for multiplexing a large number of low-rate circuits onto the high-rate fiber channel. The "basic" transmission rate of SONET is 64 kbps for supporting voice communications. SONET multiplexes large numbers of 64-kbps channels onto higher-rate datastreams. SONET defines a family of supported data rates. These data rates are often referred to as (optical carrier) OC-1, OC-3, OC-12, OC-48, and OC-192; where OC-1 corresponds to a data rate of 51.84 Mbps, or 672 voice circuits, OC-3 is 155.52 Mbps or 2016 voice circuits (3 times OC-1, OC-12 is 12 times OC-1, etc.). SONET uses time-division multiplexing (TDM) for combining traffic from multiple sources onto a common output. TDM multiplexes traffic from different sources by interleaving small "slices" of data from each source. Thus, if traffic from three OC-1 sources is being time division multiplexed onto an OC-3 transmission channel, each source would get access to the channel for a short period of time in a round-robin order, as shown in Fig. 1.

However, because of fundamental limits on optical transmission, the transmission capacity of a fiber cannot be increased indefinitely. Hence, to further increase the capacity of a fiber, a technology called wavelength-division multiplexing (WDM) was developed [1]. Wavelength division multiplexing allows transmissions on the fiber to use different colors of light (each color represents a different wavelength over which light propagates). Whereas in the first optical communications networks, light was transmitted through the fiber using a single wavelength, WDM permits light at multiple, different wavelengths, to be transmitted through a single fiber simultaneously. WDM is analogous to frequency-division multiplexing (FDM), which is often used for transmission over the airwaves.

In WDM systems, incoming optical signals are assigned a specific wavelength and then multiplexed onto the fiber. Moreover, such systems are bit-rate- and protocolindependent, meaning that each incoming signal can be carried in its native format and at a different rate. For example, a WDM system may support the transmission of multiple SONET signals on a single fiber, each operating at transmission rates of 10 Gbps (OC-192). As shown in Fig. 2, WDM systems are designed to operate in the low loss region of optical fiber, around the 1.5-µm band. Typically, wavelengths are assigned in this region with a separation of 25–100 GHz; and systems supporting anywhere from 80 to 160 wavelengths are presently being deployed.

The simplest approach to using WDM is to treat each wavelength as if it were on a separate fiber and continue



Figure 1. SONET time-division multiplexing.



Figure 2. Wavelength-division multiplexing allows the transmission on multiple wavelengths within a single fiber.

to design the network using point-to-point links, as shown in Fig. 3. With this approach the available fiber capacity would in effect be increased by a factor that equals the number of wavelengths, and the network architecture would be largely unchanged. By itself, this approach leads to significant cost savings. First, in many cases existing fiber cannot meet demand, and WDM can help alleviate the fiber exhaust problem. In addition, through the use of erbium-doped fiber amplifiers (EDFAs), as shown in Fig. 3, a number of wavelengths can be simultaneously amplified. This leads to significant cost savings when compared to pre-WDM systems where each fiber would require its own amplification. Since the cost of amplification (and or regeneration) forms a large fraction of the overall network deployment cost, these cost savings through the use of EDFAs play a large role in the commercial success of WDM systems.

As described so far, from a network perspective, WDM systems are not vastly different from any other optical transmission systems. However, in addition to pure transmission technology, a number of WDM network elements have been developed that make it possible to design networks that actually route and switch traffic in the optical domain [2]. While these optical networking functions are rather limited, they have the potential of significantly enhancing network performance. In Section 2, we describe some of the basic WDM network elements. Subsequently, in Section 3 we describe architectures for future WDM optical local-area networks (LANs), and in Section 4 we describe architectural issues in the design of all-optical WDM wide-area networks (WANs). Since all-optical networks are not likely to emerge in the near future, we devote the last section of this article to discussing future WDM-based networks that use a combination of optical and electronic processing.

2. WDM NETWORK ELEMENTS

A number of optical network elements have been developed that allow for simple optical processing of signals. The simplest such device is a "broadcast star," shown in Fig. 4. In a broadcast star, each node is connected using one input and one output fiber. The fibers are then coupled or "fused" together so that a signal coming in from any input fiber will propagate on all output fibers. In this way, all nodes can communicate with each other. Because of its simplicity and broadcast property, a star has often been proposed as a suitable technology for optical LANs. In Section 3 of this article we describe WDM-based LAN architectures utilizing a broadcast star.

A somewhat more sophisticated device is a *wavelength* router, a passive optical device that combines signals from a number of input fibers and "routes" those signals in a static manner, based on the wavelength on which they propagate, to the output fibers. The operation of a wavelength router with four input fibers is shown in Fig. 5. Notice that the signal propagating on wavelength 1 of the first input fiber is routed to the first output fiber, while the signal propagating on wavelength 2 is routed to the second output fiber, and so on. Similarly, the signal propagating on wavelength 1 of the second fiber is routed directly to the second output fiber, while the signal on wavelength 2 is routed to the third output fiber, and so forth. A wavelength router is different from a broadcast star in that it separates the wavelengths onto different output fibers. Notice that the signals propagating on the first input fiber are routed onto the four output fibers so that the first wavelength is routed to the first output, the second to the second output, and so on. In this way, a wavelength router is commonly used in optical networks as a WDM demultiplexer. WDM demultiplexers can be used to form optical add/drop multiplexers (OADMs), that are able to drop or add any number of wavelengths at a node. OADMs play an important role in a network and can be designed in a number of ways (not necessarily using a



Figure 3. Typical SONET/WDM deployment.



Figure 4. Broadcast star.



Figure 5. Wavelength router.

wavelength router); the operation of an OADM is shown in Fig. 6.

WDM demultiplexers can be used in conjunction with optical switches to form an even more sophisticated optical switching device known as a frequency-selective switch (FSS). Unlike a wavelength router that routes wavelength from input fibers onto output fibers in a static manner, a FSS is a configurable device that can take any wavelength from any input fiber and switch it onto any output fiber. In this way, a FSS allows for some flexibility in the operation of the network. The basic operation of a FSS is shown in Fig. 7. The signals traveling on each input fiber are demultiplexed into the different wavelengths. Each wavelength is then connected to a switching element that can switch any of the input fibers onto any of the output fibers. The outputs of the switch elements are then connected to WDM multiplexers that combine the signals onto the output fibers.

While a FSS adds significant functionality to the operation of the network, it requires the ability to dynamically switch optical signals from one input fiber onto another. Such switching can be accomplished optically using a number of techniques that are beyond the scope of this article. Optical switching technology, while rapidly progressing, is still relatively immature and costly. Alternatively, the signals can be switched in the electronic domain by first converting the signals traveling on each wavelength to electronics, switching the signals electronically, and retransmitting the signal onto the output fiber. Such optoelectronic switching is often less costly, and results in faster switching times, than does optical switching. However, in order to perform the optoelectronic conversion, the signal format (e.g., modulation technique, bit rate) must be known. This eliminates the signal "transparency" that is so desirable in optical networks.

A FSS adds flexibility to network operations by allowing wavelengths to be dynamically switched among the different fibers. However, notice that wavelength cannot be switched arbitrarily. For example, it is not possible for a FSS to route the signal propagating on a given wavelength from more than one input fiber onto the same output fiber. This limitation can be overcome by a wavelength converter, which can convert a signal from one wavelength onto another. Wavelength converters come in a number of variations. The simplest is a *fixed-wavelength* converter, which can convert a given wavelength onto another in a predetermined static fashion. More flexible converters can convert a given wavelength onto one of a number of wavelengths dynamically; such converters are known as *limited-wavelength converters*. The most flexible wavelength converters can convert any wavelength onto any other. Wavelength conversion can be accomplished in either the optical or electronic domain. Optical wavelength conversion is a rather immature technology primarily implemented in experimental laboratories; while electronic wavelength conversion suffers from the need for optoelectronic conversion and the consequent loss of transparency. Hence, while desirable, wavelength conversion in optical networks is still very limited.

Of course, in order to use WDM technology, one must be able to transmit and receive the signal on the different wavelengths. Transmission is accomplished using lasers that operate at a given wavelength, while reception is accomplished using WDM filters and light detectors. Typically lasers and filters are designed to operate at a single, fixed, frequency; such devices are commonly referred to as *fixed tuned devices*. Fixed tuned WDM transmitters and receivers again limit the capability and flexibility of an optical network because a signal that is transmitted on a given wavelength must travel throughout the network, and be received on that wavelength. Hence, without wavelength conversion, a node that has a transmitter that operates on a given frequency can



Figure 6. Optical add/drop multiplexer.



Figure 7. A frequency-selective switch.

communicate only with nodes that are equipped with a receiver for that frequency. Tunable transmitters and receivers are hence very desirable for optical networks; however, much like wavelength conversion, that technology is still at its inception. Tunable transmitters and receivers are often characterized according to the speed with which they can tune to different wavelengths. Slow tuning lasers, which can tune on the order of a few milliseconds, are now becoming commercially available, while fast lasers that can tune in microseconds are emerging.

3. WDM LOCAL AREA NETWORKS

Typically, local-area networks (LANs) span short distances, ranging from a few meters to a few thousands of meters. Because of the relatively close proximity of nodes, LANs are typically designed using a shared transmission medium. In this section we discuss WDM-based LANs, where users share a number of wavelengths, each operating at moderate rates (e.g., 40 wavelengths at 2.5 Gbps each).

Typically WDM-based LANs assume the use of a broadcast star architecture [3]. An optical star coupler is used to connect all the nodes. Each node is attached to the star using a pair of fibers: one for transmission and the other for reception. The star coupler is a passive device that simply connects all the incoming and outgoing fibers so that any transmission, on any wavelength, on an incoming fiber is broadcast on all outgoing fibers. In order for nodes to communicate, they must tune their transmitters and receivers to the appropriate wavelength.

A WDM LAN based on a broadcast star architecture can provide a transmission capacity that can easily exceed 100 Gbps. Perhaps the greatest reason preventing such systems from emerging is the cost of WDM transceivers. In order for a WDM LAN to allow flexible bandwidth sharing, both transmitter and receiver must be rapidly tunable over the available wavelengths. Transceiver tuning times that are smaller than the packet transmission times are desirable if efficient use of the bandwidth is to be obtained. With packets that are just a few thousands of bits in length, this calls for tuning times on the order of microseconds or faster. Present technology for fast tuning lasers is largely at the experimental stage; and while such lasers are slowly becoming commercially available, they are very expensive. Similarly, fast tuning receivers are also complex and expensive.

It is reasonable to expect that as the commercial market for these devices develops, their cost will decrease and they will become more widely available. However, in the near future, if WDM-based LANs are to become a reality they must limit the use of tunable components. WDMbased LANs are usually classified according to the number and tunability of the transmitters and receivers [4]. For example, a system utilizing one tunable transmitter and one tunable receiver is referred to as a TT-TR system. Similarly, a fixed tuned system would be referred to as FT-FR. Obviously, a FT-FR system can only use one wavelength if full connectivity among the nodes is desired. In order to provide full connectivity over multiple wavelengths, it is necessary that either the receivers or the transmitters be tunable. Systems employing either a tunable transmitter and a fixed tuned receiver (TT-FR) or a fixed transmitter and a tunable receiver (FT-TR) have been proposed in the past for the purpose of reducing the network costs.

Particularly attractive is the use of a fixed tuned receiver, because with a fixed tuned receiver all communication to a node is done on a fixed wavelength. Hence, this eliminates the need for any coordination before the transmission takes place. Of course, having a fixed tuned receiver means that nodes will have to be assigned to wavelengths in some fashion. For example, in an Nnode-W wavelengths network, N/W nodes can be assigned to receive on each wavelength. This, of course, creates a number of complications. First, when nodes are assigned to wavelengths in such a fixed manner, it is possible that certain wavelengths will be carrying a larger load than others and so, while some wavelengths may be lightly loaded, others may be overly saturated. In addition, such a network is complicated to administer because whenever adding a new node, care must be taken to determine on which wavelength it must be added, and a transceiver card tuned to that wavelength must be used.

In order to obtain the full benefit of the WDM bandwidth, a WDM-based LAN must have a TT-TR architecture. With this architecture, some form of transmission coordination is necessary for three reasons: (1) if two nodes transmit on the same wavelength simultaneously, their transmissions will interfere with each other (collide) and so some mechanism must be employed to prevent such collisions; (2) if two or more nodes transmit to the same node at the same time (albeit on different wavelengths), and if that node has only a single receiver, it will be able to receive a transmission on a wavelength, it must know in advance of the upcoming transmission so that it can tune its receiver to the appropriate wavelength.

Most proposed WDM LANs use a separate control channel for the purpose of pretransmission coordination. Often, these systems use an additional fixed tuned transceiver for the control channel. Alternatively, the control and data channels can share a transceiver, as shown in Fig. 8.



Figure 8. User terminals: (**a**) single turnable transceiver; (**b**) two-transceiver configuration.

In order for one node to send a packet to another, it must first choose a wavelength on which to transmit, and then inform the receiving node, on the control channel, of that upcoming transmission. A number of medium access control (MAC) protocols have been proposed to accomplish this exchange [5]. These protocols are more complicated than single-channel MAC protocols because they must arbitrate among a number of shared resources: the data channels, the control channel, and the receivers.

Early MAC protocols for WDM broadcast networks attempted to use ALOHA for sharing the channels [6]. With ALOHA, nodes transmit on a channel without attempting to coordinate their transmissions with any of the other nodes. If no other node transmits at the same time, the transmission is successful; however, if two nodes transmit simultaneously, their transmissions "collide" and both nodes must retransmit their packets. To reduce the likelihood of repeated collisions, nodes wait a random delay before attempting retransmission. When the load on the network is light, the likelihood of such a collision is low; however, with increased load such collisions occur more often, limiting network throughput. Single-channel versions of ALOHA have a maximum throughput of approximately 18%. A slotted version of ALOHA, where nodes are synchronized and transmit on slot boundaries, can achieve a throughput of 36%.

In a WDM system using a control channel, a MAC protocol must be used both for the control and the data channels. Early MAC protocols attempted to use a variation of ALOHA on both the control and the data channels [7]. In order for a transmission to be successful, the following sequence of events must take place: (1) the transmission on the control channel must be successful (i.e., no control channel collision), (2) the receiving node must not be receiving any other transmission at the same time (i.e., no receiver collision), and (3) the transmission on the chosen data channel must also be successful. In a system that uses ALOHA for both the control and data channels, it is clear that throughput will be very limited. It has been shown that systems using slotted ALOHA for both the data and control channels achieve a maximum utilization of less than 10% [7].

In view of the discussion above, a number of MAC protocols that attempt to increase utilization by coordinating and scheduling the transmissions more carefully have been proposed [4]. For example, the protocol described by Modiano and Barry [8] uses a simple master/slave scheduler as shown in Fig. 9. All nodes send their requests to the scheduler on a dedicated control wavelength, λ_C . The scheduler, located at the hub, schedules the requests and informs the nodes on a separate wavelength, $\lambda_{C'}$, of their turn to transmit.

Figure 9. A WDM-based LAN using a broadcast star and a scheduler.

OT

On receiving their assignments, nodes immediately tune to their assigned wavelength and transmit. Hence nodes do not need to maintain any synchronization or timing information. By measuring the amount of time that nodes take to respond to the assignments, the scheduler is able to obtain an estimate of each node's round-trip delay to the hub. This estimate is used by the scheduler to overcome the effects of propagation delays. The system uses simple scheduling algorithms that can be implemented in real time. Unicast traffic is scheduled using first-come-firstserve input queues and a window selection policy to eliminate head-of-line blocking, and multicast traffic is scheduled using a random algorithm [9].

We should point out, however, that despite their appeal, WDM-based LANs still face significant economic challenges. This is because the cost of WDM transceivers (especially tunable) is far greater than the typical cost of today's LAN interfaces. Since tunable WDM transceivers are just beginning to emerge in the marketplace, it is difficult to provide an accurate cost estimate for these devices, but it is certainly in the thousands of dollars. While a 100-Gbps LAN is very attractive, few would be willing to pay thousands of dollars for such LAN interfaces. Hence, in the near term, it is reasonable to expect WDM LANs to be used only in experimental settings or in networks requiring very high performance. However, as the cost of transceivers declines, it is not unlikely that this technology will become commercially viable.

4. WDM WIDE-AREA NETWORKS

In the WDM broadcast LAN architecture described above, nodes communicate by tuning their transmitter and receiver to common wavelengths. In a wide-area network (WAN), where a broadcast architecture is not scalable, traffic must be switched and routed at various communication nodes throughout the network. In electronic networks, this switching is accomplished using either circuit switching or packet switching techniques. With packet switching, each network node must process each packet's header to determine the destination of the packet and make suitable routing decisions; with circuit switching, circuits are set up in advance of the communication and routing and switching decisions are predetermined for the duration of the call. Hence, with circuit switching there is no need for nodes to process the incoming data.

Optical packet switching involves a number of rather complex functions, such as header recognition, packet synchronization, and optical buffering. These technologies are rather crude and largely experimental at present [10]. Hence most efforts at optical networking have focused on circuit-switched networks. With WDM, much of the effort has been on the design of wavelength-routed networks, where connections between end nodes in the network utilize a full wavelength. There are a number of challenges in the design of an optical WDM network including the choice of a network architecture, performing the functions of routing and switching wavelengths, as well as assigning wavelengths to the various connections.

Since optical network elements are relatively expensive and of limited capabilities, the choice of a network architecture is particularly critical. Early efforts at designing all-optical WDM networks have been focused on a hierarchical architecture where different network elements are employed at different levels of the hierarchy. For example, as shown in Fig. 10, an optical star may be used in the local areas of the network and frequency-selective switches, optical amplifiers, wavelength converters, and other components may be used in the backbone of the network.

An early prototype of an all-optical-network is the alloptical network (AON) testbed developed by scientists at MIT, AT&T, and Digital Equipment Corporation (DEC) [2]. The AON testbed used the hierarchical architecture shown in Fig. 11. The lowest level in the



Figure 11. The AON architecture.

hierarchy was the LAN employing a broadcast star. A number of wavelengths were allocated for use within the local area, and those were separated from the other levels of the hierarchy using a wavelength blocking filter. Separating the local wavelengths from the rest of the network allows for those wavelengths to be reused at different local areas. The next level of the hierarchy used a wavelength router for connecting (in a static manner) different local areas. Finally, a FSS was used for providing connectivity in the wide area. A prototype of the architecture consisting of the lowest two levels was deployed in the Boston area connecting between facilities at MIT, MIT Lincoln Laboratory, and DEC.

The AON testbed supported two primary types of services: (1) a circuit-switched wavelength service that can establish wavelength connectivity between different nodes and (2) a circuit-switched time-slotted service by which a fraction of a wavelength can be assigned to a connection. The time-slotted service allows the flexibility of provisioning at the subwavelength level. However, implementing such a service requires very precise synchronization between the nodes so that the different circuits can be aligned on time-slot boundaries. The AON testbed demonstrated a 20-wavelength network, separated by 50 GHz and transmitting at rates of up to 10 Gbps per wavelength. AON also employed tunable transceivers. The transmitter was implemented using a DBR (distributed Bragg reflector) laser that can tune between wavelengths in 10 ns.

The early wavelength routing networks raised a number of architectural questions for all-optical networks. Perhaps the one that received the most attention is that of dealing with wavelength conflicts. Without using a wavelength converter, the same wavelength must be available for use on all links between the source and the destination of the call. A wavelength conflict may occur when each link on the route may have some free wavelengths, but the same wavelength is not available on all of the links. This situation can be dealt with through the use of wavelength converters that can switch between the wavelengths. However, because of the high cost of wavelength conversion, a number of studies quantifying the benefits of wavelength conversion in a network have been published [11,12]. Others have considered the possibility of placing the wavelength converters only at some key nodes in the networks [13]. However, the detailed results of these studies are beyond the scope of this article.

Another promising approach for dealing with wavelength conflicts is the use of a good wavelength assignment algorithm that attempts to reduce the likelihood of a wavelength conflict occurring. The wavelength assignment algorithm is responsible for selecting a suitable wavelength among the many possible choices for establishing the call. For example, the three calls illustrated in Fig. 12 can be established using three wavelengths $(\lambda_1\lambda_2\lambda_3)$ as shown on the left or just two wavelengths as shown on the right. By choosing the assignment on the right, λ_3 remains free for use by future potential calls. A number of wavelength assignment schemes have been proposed [14,15], and the subject remains an active area of research.



Figure 12. Two possible wavelength assignments for three calls on a ring: (**a**) bad and (**b**) good assignments.



Figure 13. Performance of wavelength assignment algorithms in a ring network.

Figure 13 compares the performance of some proposed wavelength assignment algorithms. The simplest algorithm is to randomly select a wavelength from among the available wavelengths along the path. Clearly such an algorithm would be very inefficient, and, as can be seen from the figure, the random algorithm results in the highest blocking probability. A first-fit heuristic assigns the first available (i.e., lowest index number) wavelength that can accommodate the call. The most frequently used heuristic assigns the wavelength that is used on the most number of fibers in the network and lastly, the maxsum algorithm assigns the wavelength that maximizes the number of paths that can be supported in the network after the wavelength has been assigned [16]. All of these algorithms attempt to pack the wavelengths as much as possible, leaving free wavelengths open for future calls. Also shown in Fig. 12 is the blocking probability that results when wavelength changers are used. This represents an upper bound on the performance of any wavelength assignment algorithm. The significance of this illustration is that a good wavelength assignment algorithm can result in a blocking probability that is nearly as low as if wavelength changers are employed. Hence, a significant reduction in network costs can be obtained by using a good wavelength assignment algorithm.

Wavelength assignment was the first fundamental architectural problem in the design of all-optical networks. It has received much attention in the literature, and remains an active area of research. Beyond wavelength assignment, other important areas of research include the use of wavelength conversion (e.g., where wavelength converters should be deployed), the use of optical switching, and mechanisms for providing protection from failures. While a meaningful discussion of these topics is beyond the scope of this article, Ref. 1 provides a recent overview on most of these topics.

5. JOINT OPTICAL AND ELECTRONIC NETWORKS

Since all-optical networks are not likely to become a reality in the near future, the current trend in networking is to design networks that use a combination of optical and electronic techniques. A simple example would be a SONET-over-WDM network where the nodes in a SONET ring are connected via wavelengths rather than point-topoint fiber links. As we explained in the introduction, this use of WDM transmission is beneficial because it both reduces network cost and increases network capacity due to the large number of wavelengths. In fact, using WDM at the optical layer also introduces an additional flexibility in the design of the network.

Consider, for example, the networks in Fig. 14, where the optical topology consists of optical nodes [e.g., optical switches or ADMs (add/drop multiplexes)] that are connected via fiber and the electronic topology consists of electronic nodes (e.g., SONET multiplexers) that are connected using electronic links. Without WDM, the electronic topology shown in the figure cannot possibly be realized on the optical topology because the optical topology does not have a fiber link between nodes 1 and 3. However, with WDM an electronic link can be established between nodes 1 and 3 using a wavelength that is routed through node 2. The optical switch (or ADM) at node 2 can be configured to pass that wavelength through to node 3, creating a virtual link between nodes 1 and 3. This approach allows for various electronic topologies to be realized on optical topologies that do not necessarily have the same structure. Electronic nodes can be connected via wavelengths that are routed on the optical topology.

This approach can be used to realize a variety of electronic networks, such as ATM, IP, or SONET [17]. The connectivity of the electronic nodes determines the required wavelength connection that must be established. In other words, each link in the electronic topology requires a wavelength connection (also referred to as *lightpath*) between the optical nodes. In order to realize a particular electronic topology, the corresponding set of wavelength connections must be realized on the optical topology. This very practical problem leads to another version of the routing and wavelength assignment (RWA) problem known as the *batch RWA*. Given a set of lightpaths that must be established, a RWA must be found such that each lightpath must use the same wavelengths along its



Figure 14. The electronic (a) and optical (b) topologies of a network.

route from the source to the destination (assuming no wavelength conversion) and no two lightpaths can use the same wavelength on a given link. This problem is closely related to the well-known NP-complete graph coloring problem, and in fact Chlamtac et al. [17] showed that the static RWA problem is indeed NP-complete by suitable transformation from graph coloring.

WDM allows the electronic (logical) topology to be different from the physical topology over which it is implemented. This ability created the interesting opportunity for "logical topology design"; that is, given the traffic demand between the different nodes in the network, what is the best logical topology for supporting that demand. For example, suppose that one is to implement a ring logical topology as in Fig. 14. If a large amount of traffic is being carried between nodes 1 and 4, and virtually no traffic between 1 and 3, it makes more sense to connect the ring in the order 1-4-3 rather than 1-3-4. In this way, the length of the path that the traffic must traverse is reduced, and consequently, the load on the links is also reduced. Designing logical topologies for WDM networks has typically been formulated as an Integer Programming problem, solutions to which are obtained using a variety of search heuristics [18,19]. Furthermore, with configurable WDM nodes (e.g., wavelength switches), it is even possible to reconfigure the logical topology in response to changes in traffic conditions [20].

6. FUTURE DIRECTIONS

Since their inception, the premise of optical networks has been to eliminate the electronic bottleneck. Yet, so far, all optical networks have failed to emerge as a viable alternative to electronic networks. This can be attributed to the tremendous increase in the processing speeds and capacity of electronic switches and routers. Nonetheless, the enormous capacity and configurability of WDM can be used to reduce the cost and complexity of the electronic network, paving the way to even faster networks.

While in the near term optical networking will be used only as a physical layer beneath the electronic network, researchers are aggressively pursuing all-optical packetswitched networks. Such networks may first emerge in the local area, where a broadcast architecture can be used without the need for optical packet switching. As alloptical packet switching technology matures, all-optical networks may become viable. However, as of this writing, it is very unclear whether truly all-optical networks will ever become a reality.

BIOGRAPHY

Eytan Modiano received his B.S. degree in electrical engineering and computer science from the University of Connecticut at Storrs in 1986 and his M.S. and Ph.D. degrees, both in electrical engineering, from the University of Maryland, College Park, Maryland, in 1989 and 1992, respectively. He was a Naval Research Laboratory fellow between 1987 and 1992 and a National Research Council postdoctoral fellow during 1992–1993, while he was

2846 WAVELETS: A MULTISCALE ANALYSIS TOOL

conducting research on security and performance issues in distributed network protocols.

Between 1993 and 1999 he was with the Communications Division at MIT Lincoln Laboratory where he designed communication protocols for satellite, wireless, and optical networks and was the project leader for MIT Lincoln Laboratory's Next Generation Internet (NGI) project. Since 1999, he has been a member of the faculty of the Aeronautics and Astronautics Department and the Laboratory for Information and Decision Systems (LIDS) at MIT, where he conducts research on communication networks and protocols with emphasis on satellite and hybrid networks, and high speed optical networks.

BIBLIOGRAPHY

- R. Ramaswami and K. Sivarajan, *Optical Networks*, Morgan Kaufmann, San Francisco, CA, 1998.
- I. P. Kaminow, A wideband all-optical WDM network, *IEEE JSAC/JLT* 14(5): 780–799 (June 1996).
- E. Modiano, WDM-based packet networks, *IEEE Commun.* Mag. 37(3): 130-135 (March 2000).
- B. Mukherjee, WDM-based local lightwave networks part I: Single-hop systems, *IEEE Network* 6(3): 12-27 (May 1992).
- G. N. M. Sudhakar, M. Kavehrad, and N. D. Georganas, Access Protocols for passive optical star networks, *Comput. Networks ISDN Syst.* 913–930 (1994).
- D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
- N. Mehravari, Performance and protocol improvements for very high-speed optical local area networks using a star topology, *IEEE/OSA J. Lightwave Technol.* 8(4): 520-530 (April 1990).
- E. Modiano and R. Barry, A novel medium access control protocol for WDM-based LAN's and access networks using a master/slave scheduler, *IEEE J. Lightwave Technol.* 18(4): 2-12 (April 2000).
- E. Modiano, Random algorithms for scheduling multicast traffic in WDM broadcast-and-select networks, *IEEE/ACM Trans. Network.* 7(3): 425-434 (June 1999).
- V. Chan, K. Hall, E. Modiano, and K. Rauchenbach, Architectures and technologies for optical data networks, *J. Lightwave Technol.* 16(12): 2146–2186 (December 1998).
- R. Barry and P. Humblet, Models of blocking probability in all-optical networks with and without wavelength changers, *IEEE J. Select. Areas Commun.* 14(5): 858–867 (June 1996).
- S. Subramanian, M. Azizoglu, and A. Somani, Connectivity and sparse wavelength conversion in wavelength-routing networks, *IEEE INFOCOM* 148-155 (1996).
- R. Ramaswami and G. Sasaki, Multiwavelength optical networks with limited wavelength conversion, *IEEE INFOCOM* 490-499 (1997).
- R. Ramaswami and K. N. Sivarajan, Routing and wavelength assignment in all-optical networks, *IEEE/ACM Trans. Network*. 489-500 (Oct. 1995).
- 15. S. Subramanian and R. Barry, Wavelength assignment in fixed routing wdm networks, ICC '97, Montreal, June 1997.
- I. Chlamtac, A. Ganz, and G. Karmi, Lightpath communications: An approach to high-bandwidth optical WANs, *IEEE Trans. Commun.* 40(7): 1171–1182 (July 1992).

- J. Bannister, L. Fratta, and M. Gerla, Topological design of WDM networks, *Proc. Infocom* '90, Los Alamos, CA, 1990.
- R. Ramaswami and K. Sivarajan, Design of logical topologies for wavelength routed optical networks, *IEEE J. Select. Areas Commun.* 14(5): 840–851 (June 1996).
- J. P. Labourdette and A. Acampora, Logically rearrangeble multihop lightwave networks, *IEEE Trans. Commun.* 39(8): 1223-1230 (Aug. 1991).
- A. Narula-Tam and E. Modiano, Dynamic load balancing in WDM packet networks with and without wavelength constraints, *IEEE J. Select. Areas Commun.* 18(10): 1972–1979 (Oct. 2000).

WAVELETS: A MULTISCALE ANALYSIS TOOL

HAMID KRIM ECE Department, North Carolina State University Centennial Campus Raleigh, North Carolina

One's first and most natural reflex when presented with an unfamiliar object is to carefully look it over, and hold it up to the light to inspect its different facets in the hope of recognizing it, or of at least relating any of its aspects to a more familiar and well-known entity. This almost innate strategy pervades all science and engineering disciplines.

Physical phenomena (e.g., earth vibrations) are monitored and observed by way of measurement in the form of temporal and/or spatial data sequences. Analyzing such data is tantamount to extracting information useful to further understand the underlying process (e.g., frequency and amplitude of vibrations may be an indicator for an imminent earthquake). Visual or manual analysis of typically massive amounts of acquired data (e.g., in remote sensing) are impractical, causing one to resort to adapted mathematical tools and analysis techniques to better cope with potential intricacies and complex structure of the data. Among these tools figure a variety of functional transforms (e.g., Fourier transform) that in many cases may facilitate and simplify an analytical track of a problem, and frequently (and just as importantly) provide an alternative view of, and a better insight into, the problem. (This, in some sense, is analogous to exploring and inspecting data under a "different light.") An illustration of such a "simplification" is shown in Fig. 1, where a rather intricate signal x(t) shown in the leftmost figure may be displayed or viewed in a different space as two elementary tones. In Fig. 2, a real bird chirp is similarly displayed as a fairly rich signal which, when considered in an appropriate space, is reduced and "summarized" to a few "atoms" in the time-frequency (TF) representation. Transformed signals may formally be viewed as convenient representations in a different domain that is itself described by a set of vectors/functions $\{\phi_i(t)\}_{i=\{1,2,\dots,N\}}$. A contribution of a signal x(t) along a direction " $\phi_i(t)$ " (its projection) is given by the following inner product

$$C_i(x) = \langle x(t), \phi_i(t) \rangle = \int_{-\infty}^{\infty} x(t)\phi_i(t) dt$$
 (1)





Figure 2. Bird chirp with a simplified representation in an appropriate basis.

where the compact notation $\langle \cdot, \cdot \rangle$ for an inner product is used. The choice of functions in the set is usually intimately tied to the nature of information that we are to extract from the signal of interest. A Fourier function for example, is specified by a complex exponential " $e^{j\omega t}$ " and reflects the harmonic nature of a signal, and provides it with a simple and an intuitively appealing interpretation as a "tone." Its spectral (frequency) representation is a Dirac impulse that is well localized in frequency. A Fourier function is hence well adapted to represent a class of signals with a fairly well localized "spectrum," and is highly inadequate for a transient signal that, in

Figure 1. A canonical function-based projection to simplify a representation.

contrast, tends to be temporally well localized. This thus calls for a different class of analyzing functions, namely, those with good temporal compactness. A relatively recent introduction of a function class that balances between the two extremes will be the focus of this chapter. The wavelet transform, by virtue of its mathematical properties, indeed provides such an analysis framework that in many ways is complementary to that of the Fourier transform. As the goal of this article is of a tutorial nature, we cannot help but provide a somewhat high-level exposition of this theoretical framework while attempting to provide a working knowledge of the tool as well as its potential for applications in information sciences in general.

The remainder of this article is organized as follows. In Section 1 we review some of the background starting with a generic functional basis with illustrations from the familiar Fourier basis. In Section 2 we discuss the wavelet transform. In Section 3 we define a multiscale analysis based on a wavelet basis and elaborate on their properties and their implications, as well as on their applications. In Section 4 we discuss a specific estimation technique that is believed to be sufficiently generic and general to be useful in a number of different applications of information sciences in general and of signal processing and communications in particular.

1. SIGNAL REPRESENTATION IN FUNCTIONAL BASES

As noted above, a proper selection of a set of analyzing functions heavily impacts the resulting representation of an observed signal in the dual space, and hence the resulting insight as well. When considering finite-energy signals [these are also said to be $L^2(\mathbb{R})$]

$$\int_{-\infty}^{\infty} |x(t)|^2 dt < \infty, \tag{2}$$

a convenient functional vector space is one that is endowed with a norm inducing inner product, namely, a Hilbert space, which will also be our assumed space throughout.

Definition 1. A vector space endowed with an inner product, which in turn induces a norm, and such that every sequence of functions $x_n(t)$ whose elements are asymptotically close is convergent, (this convergence is in the Cauchy sense whose technical details are deliberately omitted to streamline the flow of the article) is called a Hilbert space.

Remark. An $L^2(\mathbb{R}^d)$, d = 1, 2 space is a Hilbert space.

Definition 2. A set of vectors or functions $\{\phi_i(t)\}, i = 1, ..., which are linearly independent and which span a vector space, is called a basis. If in addition these elements are orthonormal, then the basis is said to be an orthonormal basis.$

Among the desirable properties of a selected basis in such a space are *adaptivity* and a *capacity* to preserve and reflect some key signal characteristics (e.g., signal smoothness). Fourier bases have in the past been at the center of all science and engineering applications. They have in addition, provided an efficient framework for analyzing periodic as well as nonperiodic signals with dominant modes.

1.1. Fourier Bases

If a canonical function of a selected basis is $\phi(t) = e^{j\omega t}$, where $\omega \in \mathbb{R}$, then we speak of a Fourier basis that yields a Fourier series (FS) for periodic signals and a Fourier transform (FT) for aperiodic signals.

A FT essentially measures the spectral content of a signal x(t) across all frequencies ω :

$$\hat{X}(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt.$$
(3)

The latter, also referred to as the *Fourier integral*, is defined for a broad class of signals [1], and may also be specialized to derive the Fourier series (FS) of a periodic signal. This is easily achieved by noting that a periodic signal $\tilde{x}(t)$ may be evaluated over a period T, which in turn leads to

$$\tilde{x}(t) = x(t+T) = \sum_{n=-\infty}^{\infty} \alpha_n^x e^{jn\omega_0 t}$$
(4)

with $\omega_0 = 2\pi/T$ and $\alpha_n^x = 1/T \int_0^T x(t)e^{-jn\omega_0 t} dt$. Because it is an orthonormal transform, the FT enjoys a number of interesting properties [2], including the energy preservation in the dual space (transform domain):

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{X}(f)|^2 df.$$

This is referred to as the Plancherel-Parseval property [2].

In applications, however, x(t) is measured and sampled at discrete times, requiring that the aforementioned transform be extended to obtain a spectral representation that closely approximates the theoretical truth and that remains just as informative. Toward that end, we proceed to define a discrete-time Fourier transform (DFT) as

$$\hat{X}(e^{j\omega}) = \sum_{i=-\infty}^{\infty} x(i)e^{-j\omega i}$$
(5)

This expression may also be extended for finiteobservation (finite-dimensional) signals via the fast Fourier transform [2]. While these transforms have been and remain crucial in many applications, they show limitations in problems requiring a good "time-frequency" localization as often encountered in transient analysis. This may easily be understood by reinterpreting Eq. (5) as a weighted transform where each of the time samples is equally weighted in the summation for $\hat{X}(e^{j\omega})$. Gabor [3] first proposed to use a different weighting window leading to the so-called windowed FT, which served well in many practical applications [4].

1.2. Windowed Fourier Transform

As just mentioned, when our goal is to analyze very local features, such as those present in transient signals, for instance, it then makes sense to introduce a focusing window as follows

$$W_{\mu,\omega}(t) = e^{j\omega t} W(t-\mu),$$

where $||W_{(\mu,\omega)}|| = 1 \forall (\mu, \omega) \in \mathbb{R}^2$ and where $W(\cdot)$ is typically a smooth function with a compact support. This yields the following parameterized transform:

$$\hat{X}_{W}(\omega,\mu) = \langle x(t), W_{\mu,\omega}(t) \rangle.$$
(6)

The selection of a proper window is problem-dependent and is ultimately resolved by the desired spectrotemporal tradeoff which is itself constrained by the Heisenberg uncertainty principle [4,5]. From the TF (time-frequency) distribution perspective, and as discussed by Gabor [3] and displayed in Fig. 3, the Gaussian window may be shown to have minimal temporal as well as spectral support of any other function. It hence represents the best compromise between temporal and spectral resolutions. Its numerical implementation entails a discretization of the modulation and translation parameters and results in a uniform partitioning of the TF plane as illustrated in Fig. 4. Different windows result in various TF distribution of elementary atoms, favoring either temporal or spectral resolution as may be seen for the different windows in Fig. 5. While representing an optimal time-frequency compromise, the uniform coverage of the TF plane by the Gabor transform falls short of adequately resolving a signal whose components are spectrally far apart. This may easily and convincingly be illustrated by the study case in Fig. 6, where we note the number of cycles that may be enumerated within a window of fixed time width. It is readily



Figure 3. A Gaussian waveform results in a uniform analysis in the time-frequency plane.



Figure 4. A time-frequency tiling of dyadic wavelet basis by a proper subsampling of a wavelet frame.



Figure 5. Tradeoff resulting from windows.



Figure 6. Time windows and frequency tradeoff.

seen that while the selected window (shown grid) may be adequate for one fixed frequency component, it is inadequate for another lower frequency component. An analysis of a spectrum exhibiting such a time-varying behavior is ideally performed by way of a frequency-dependent time window as we elaborate in the next section. The wavelet transform described next, offers a highly adaptive window that is of compact support, and that, by virtue of its dilations and translations, covers different spectral bands at all instants of time.

2. WAVELET TRANSFORM

Much like the FT, the WT is based on an elementary function, which is well localized in time and frequency. In addition to a compactness property, a function has to satisfy a set of properties to be admissible as a wavelet. The first fundamental property is stated next.

Definition 3. A wavelet is a finite-energy function $\psi(\cdot)$ [*i.e.*, $\psi(\cdot) \in L^2(\mathbb{R})$] with zero mean [4,6,7]:

$$\int_{-\infty}^{\infty} \psi(t) \, dt = 0. \tag{7}$$

Commonly normalized so that $\|\psi\| = \int |\psi|^2 dt = 1$, it also constitutes a fundamental building block in the construction of functions (atoms) spanning the time-frequency plane by way of dilation and translation parameters. We hence write

$$\psi_{\mu,\xi}(t) = \frac{1}{\sqrt{\xi}}\psi\left(\frac{t-\mu}{\xi}\right)$$

where the scaling factor $\xi^{1/2}$ ensures an energy invariance of $\psi_{\mu,\xi}(t)$ over all dilations $\xi \in \mathbb{R}^+$ and translations $\mu \in \mathbb{R}$. With such a function in hand, and toward mitigating the limitation of a windowed FT, we proceed to define a



Figure 7. Admissible Mexican hat wavelet: (a) Mexihat wavelet; (b) spectrum of Mexihat.

Wavelet transform (WT) with such a capacity. A scaledependent window with good time localization properties as shown in Fig. 7 yields a transform for x(t) given by

$$\mathcal{W}_x(\mu,\xi) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{\xi}} \psi^*\left(\frac{t-\mu}{\xi}\right) dt \tag{8}$$

where the asterisk denotes complex conjugate. This is, of course, in contrast to the Gabor transform whose window width remains constant throughout. A time-frequency plot of a continuous wavelet transform is shown in Fig. 8 for a corresponding x(t).

2.1. Inverting the Wavelet Transform

Similar to the weighted FT, the WT is a redundant representation, which, with a proper normalization factor, leads to a reconstruction formula

$$x(t) = \frac{1}{C_{\psi}} \int_0^\infty \int_{-\infty}^\infty \mathcal{W}_x(\mu,\xi) \frac{1}{\sqrt{\xi}} \psi\left(\frac{t-\mu}{\xi}\right) \frac{d\xi}{\xi^2} d\mu \qquad (9)$$

with
$$C_{\psi} = \int_{0}^{+\infty} [\psi(\hat{\omega})/\omega] \, d\omega.$$

While the direct and inverse WT have been successfully applied in a number of different problems [4], their computational cost due to their continuous and redundant nature is considered as a serious drawback. An immediate and natural question arises about a potential reduction in the computational complexity and hence in the WT redundancy. Clearly, this has to be carefully carried out to guarantee a sufficient coverage of the TF plane and thereby ensure a proper reconstruction of a transformed signal. On discretizing the scale and dilation parameters, a dimension reduction relative to a continuous transform is achieved. The desired adaptivity of the window width with varying frequency naturally results by selecting a geometric variation of the dilation parameter $\xi = \xi_0^m, m \in \mathbb{Z}$ (set of all positive and negative integers). To obtain a systematic and consistent coverage of the TF plane, our choice of the translation parameter μ should be in congruence with the fact that at any scale m, the coverage of the whole line $\mathbb R$ (e.g.) is complete, and the translation parameter be in step with the chosen wavelet $\psi(t)$, that is, $\mu = n\mu_0\xi_0^m$ with $n \in \mathbb{Z}$. This hence gives the following scale and translation adaptive wavelet

$$\psi(t)_{m,n} = \xi_0^{-m/2} \psi\left(\frac{t}{\xi_0^m} - n\mu_0\right)$$
(10)

where the factor $\xi_0^{-m/2}$ ensures a unit energy function. This reduction in dimensionality yields a redundant discrete wavelet transform endowed with a structure that lends itself to an iterative and fast inversion or reconstruction of a transformed signal x(t). The set of resulting wavelet coefficients $\{< x(t), \psi_{mn}(t) >\}_{(m,n)\in\mathbb{Z}^2}$ completely characterizes x(t), and hence leads to a stable reconstruction [4,5] if $\forall x(t) \in L^2(\mathbb{R})$ (or finite energy signal) the following condition on the energy holds:

ų

$$A\|x(t)\|^{2} \leq \sum_{m,n} |\langle x(t), \psi_{mn}(t) \rangle|^{2} \leq B\|x(t)\|^{2}.$$
 (11)

Such a set of functions $\{\psi_{mn}(t)\}_{(m,n)\in\mathbb{Z}^2}$ then constitutes a frame. The energy inequality condition intuitively suggests that the redundancy should be controlled to avoid



Figure 8. Continuous signal with corresponding wavelet transform with cones of influence around singularities.

instabilities in the reconstruction (i.e., too much redundancy makes it more likely that any perturbation of coefficients will yield an unstable/inconsistent reconstruction). If the frame bounds A and B are equal, the corresponding frame is said to be tight, and if furthermore A = B = 1, it is an orthonormal basis. Note, however, that for any $A \neq 1$ a frame is not a basis. In some cases, the frame bounds specify a redundancy ratio that may help guide one in inverting a frame representation of a signal, since a unique inverse does not exist. An efficient computation of an inverse (for reconstruction), or more precisely of a pseudoinverse, may only be obtained by a judicious manipulation of the reconstruction formula [4,5]. This is, in a finite-dimensional setting, similar to a linear system of equations with a rank-deficient matrix whose column space has an orthogonal complement (the union of the two subspaces yields all the Hilbert space), and hence whose inversion is ill conditioned. The size of this space is determined by the order of the deficiency (number of linearly dependent columns), and explains the potential for instability in a frame projection-based signal reconstruction [4]. This type of "effective rank" utilization is encountered in numerical linear algebra applications (e.g., signal subspace methods [8], model order identification,). The close connection between the redundancy of a frame and the rank deficiency of a matrix suggests that solutions available in linear algebra [9] may provide insight in solving our frame-based reconstruction problem. A well-known iterative solution to a linear system and based on a gradient search solution has been described [10] (and in fact coincides with a popular iterative solution to the frame algorithm for reconstructing a signal) for solving

$$\mathbf{f} = \mathbf{L}^{-1}\mathbf{b}$$

where **L** is a matrix operating on **f** and **b** is the data vector. Starting with an initial guess \mathbf{f}_0 and iterating on it leads to

$$\mathbf{f}_n = \mathbf{f}_{n-1} + \alpha (\mathbf{b} - \mathbf{L}\mathbf{f}_{n-1}) \tag{12}$$

When α is appropriately selected, the latter iteration may be shown to yield [4]

$$\lim_{n \to +\infty} \mathbf{f}_n = \mathbf{f}.$$
 (13)

Numerous other good solutions have also been proposed and described in detail in the literature [4,5,11,12].

3. WAVELETS AND MULTIRESOLUTION ANALYSIS

While a frame representation of a signal is of lower dimension than that of its continuous counterpart, a complete elimination of redundancy is possible only by orthonormalizing a basis. A proper construction of such a basis ensures orthogonality within scales as well as across scales. The existence of such a basis with a dyadic scale progression was first shown, and an explicit construction was given by Meyer and Daubechies [5,13]. The connection to subband coding discovered by Mallat [14] resulted in a numerically stable and efficient implementation that helped propel wavelet analysis at the forefront of computational sciences. (See Daubechies and Meyer [15] for a comprehensive historical as well as technical development [5], which also led to Daubechies' celebrated orthonormal wavelet bases.)

To help maintain the smooth flow of this article and achieve the goal of endowing an advanced undergraduate with a working knowledge of the multiresolution analysis (MRA) framework, we give a high-level, albeit comprehensive, introduction, and defer most of the technical details to the sources in the bibliography. For example, the tradeoff in time-frequency resolution advocated earlier lies at the heart of the often technical wavelet design and construction. The balance between the spectral and temporal decay, which constitutes one of the key design criteria, has led to a wealth of new functional bases, and this follows the introduction of the now classical multiresolution analysis framework [4,16]. The pursuit of a more refined analysis arsenal resulted in the nonlinear multiresolution analysis introduced in 1997 and 1998 [17–19].

3.1. Multiresolution Analysis

The MRA theory developed by Mallat and Meyer [13,20], may be viewed as a functional analytic approach to subband signal analysis, which had previously been introduced in and applied to engineering problems [21,22]. The clever connection between an orthonormal wavelet decomposition of a signal and its filtering by a bank of corresponding filters led to an aximonatic formalization of the theory and subsequent equivalence. This as a result, opened up a new wide avenue of research in the development and construction of new functional bases as well as filter banks [7,23-25]. The construction of a telescopic set of nested approximation and detail subspaces $\{V_j\}_{j\in \mathbb{Z}}$ and $\{W_j\}_{j\in \mathbb{Z}}$ each endowed with an orthonormal basis, as shown in Fig. 9, is a key step of the analysis. An inter- and intrascale orthogonality of the wavelet functions, as noted above, is preserved, with the interscale orthogonality expressed as

$$W_i \perp W_j \,\forall i \neq j. \tag{14}$$



Nested wavelet subspaces and corresponding bases

Figure 9. Hierarchy of wavelet bases.

By replacing the discrete parameter wavelet $\psi_{ii}(t)$ where $(i, j) \in \mathbb{Z}^2$ (set of all positive and negative 2-tuple integers) respectively denote the translation and the scale parameters, in Eq. (8) (i.e., $\mu = 2^{j}i, \xi = 2^{J}$) we obtain the orthonormal wavelet coefficients as

$$C_{i}^{i}(x) = \langle x(t), \psi_{ij}(t) \rangle.$$
(15)

The orthogonal complementarity of the scaling subspace and that of the residuals and details amount to synthesizing the following higher-resolution subspace:

$$V_i \oplus W_i = V_{i-1}$$
.

Iterating this property, leads to a reconstruction of the original space where the observed signal lies and that, in practice, is taken to be V_0 : the observed signal at its first and finest resolution. (This may be viewed as implicitly accepting the samples of a given signal as the coefficients in an approximation space with a scaling function corresponding to that on which the subsequent analysis is based.) The dyadic scale progression has been thoroughly investigated, and its wide acceptance and popularity is due to its tight connection with subband coding whose practical implementation is fast and simple. Other nondyadic developments have also been explored [e.g., 7].

The qualitative characteristics of the MRA we have thus far discussed may be succinctly stated as follows.

Definition 4. A sequence $\{V_i\}_{i \in \mathbb{Z}}$ of closed subspaces of $L^{2}(\mathbb{R})$ is a multiresolution approximation if the following properties hold [4]:

- $\forall (j,k) \in \mathbb{Z}^2, x(t) \in V_j \leftrightarrow x(t-2^jk) \in V_j$ $\forall j \in \mathbb{Z}, V_{j+1} \subset V_j$
- $\forall j \in \mathbb{Z}, x(t) \in V_j \leftrightarrow x\left(rac{t}{2}
 ight) \in V_{j+1}$

•
$$\lim_{j \to -\infty} V_j = \bigcap_{j = -\infty} V_j = \{0\}$$

•
$$\lim_{j = -\infty} = \overline{\bigcup_{j = -\infty}^{\infty} V_j} = L^2(\mathbb{R})$$

• There exists a function $\phi(t)$ such that $\{\phi(t-n)\}_{n\in\mathbb{Z}}$ is a Riesz basis of V_0 , where the overbar denotes closure of the space.

3.2. Properties of Wavelets

A wavelet analysis of a signal assumes a judiciously preselected wavelet and hence a prior knowledge about the signal itself. As stated earlier, the properties of an analyzing wavelet have a direct impact on the resulting multiscale signal representation. Carrying out a useful and meaningful analysis is hence facilitated by a good understanding of some fundamental wavelet properties.

3.2.1. Vanishing Moments. Recall that one of the fundamental admissibility conditions of a wavelet is that its first moment be zero. This is intuitively understood in the sense that a wavelet focuses on residuals or oscillating features of a signal. This property may in fact be further exploited by constructing a wavelet with an arbitrary number of vanishing moments. We say that a wavelet has *n* vanishing moments if

$$\int \psi(t)t^{i} dt = 0, i = \{0, 1, \dots, n-1\}$$
(16)

Reflecting a bit on the properties of a Fourier Transform of a function [2], it is easy to note that the number of zero moments of a wavelet reflects the behavior of its Fourier Transform around zero. This property is also useful in applications such as compression, where it is highly desirable to maximize the number of small or negligible coefficients, and preserve only a minimal number of large coefficients. The associated cost with increasing the number of vanishing moments is that of an increased support size for the wavelet, hence that of the corresponding filter [4,5], and hence of some of its localizing potential.

3.2.2. Regularity and Smoothness. The smoothness of a wavelet $\psi(t)$ is important for an accurate and parsimonious signal approximation. For a large class of wavelets (those relevant to applications), the smoothness (or regularity) property of a wavelet, which may also be measured by its degree of differentiability $(d^{\alpha}\psi(t)/dt^{\alpha})$ or equivalently by its "Lipschitzity" γ , is also reflected by its number of vanishing moments [4]. The larger the number of vanishing moments, the smoother the function. In applications, such as image coding, a smooth analyzing wavelet is useful for not only compressing the image but for controlling the visual distortion due to errors as well. The associated cost (i.e., some tradeoff is in order) is again a size increase in the wavelet support, which may in turn make it more difficult to capture local features, such as important transient phenomena.

3.2.3. Wavelet Symmetry. At the exception of a Haar wavelet, compactly supported real wavelet are asymmetric around their centerpoint. A symmetric wavelet clearly corresponds to a symmetric filter that is characterized by a linear phase. The symmetry property is important for certain applications where symmetric features are crucial (e.g., symmetric error in image coding is better perceived). In many applications, however, it is generally viewed as a property secondary to those described above. When such a property is desired, truly symmetric biorthogonal wavelets have been proposed and constructed [5,26] with the slight disadvantage of using different analysis and synthesis mother wavelets. In Fig. 10, we show some illustrative cases of symmetric wavelets. Other nearly symmetric functions (referred to as *symlets*) have also been proposed, and a detailed discussion of the pros and cons of each is deferred to Daubechies [5].

3.3. A Filter Bank View: Implementation

As noted earlier, the connection between a MRA of a signal and its processing by a filter bank was not only of intellectual interest, but was of breakthrough proportion for general applications as well. It indeed provided a highly



Figure 10. Biorthogonal wavelets preserve symmetry and symlets nearly do: (a) symmlet; (b) symmlet spectrum; (c) spline biorthogonal wavelet; (d) spectrum of spline wavelet.



Figure 11. Filter bank implementation of a wavelet decomposition.

efficient numerical implementation for a theoretically very powerful methodology. Such a connection is most easily established by invoking the nestedness property of MR subspaces. Specifically, it implies that if $\phi(2^{-j}t) \in V_j$ and $V_j \subset V_{j-1}$, we can hence write

$$\frac{1}{2^{j/2}}\phi(2^{-j}t) = \frac{1}{2^{(j-1)/2}}\sum_{k=-\infty}^{\infty}h(k)\phi(2^{-j+1}t-k)$$
(17)

where $h(k) = \langle \phi(2^{-j}t), \phi(2^{-j+1}t - k) \rangle$, that is, the expansion coefficient at time shift k. By taking the FT of Eq. (17), we obtain

$$\Phi(2^{j}\omega) = \frac{1}{2^{1/2}} H(2^{j-1}\omega) \Phi(2^{j-1}\omega)$$
(18)

which, when iterated through scales, leads to

$$\Phi(\omega) = \prod_{p=-\infty}^{\infty} \frac{h(2^{-p}\omega)}{\sqrt{2}} \Phi(0).$$
(19)

The complementarity of scaling and detail subspaces noted earlier, stipulates that any function in subspace W_j may also be expressed in terms of $V_{j-1} = \text{Span}\{\phi_{j-1,k}(t)\}_{(j,k)\in\mathbb{Z}^2}$, or

$$\frac{1}{2^{j/2}}\psi(2^{-j}t) = \sum_{i=-\infty}^{\infty} \frac{1}{2^{(j-1)/2}} g(i)\phi(2^{-j+1}t-i).$$
(20)

In the Fourier domain, this leads to

$$\Psi(2^{j}\omega) = \frac{1}{\sqrt{2}}G(2^{j-1}\omega)\Phi(2^{j-1}\omega)$$
(21)

whose iteration also leads to an expression of the wavelet FT in terms of the transfer function $G(\omega)$, as given for $\Phi(\omega)$ in Eq. (19). In light of these equations, it is clear that the filters $\{G(\omega).$ This is illustrated for Daubechies wavelet in Fig. 12. $H(\omega)\}$ may be used to compute functions at successive scales. This in turn implies that the coefficients of any function x(t) may be similarly obtained as they are merely the result of an inner product of the signal of interest x(t) with a basis function:

$$\begin{aligned} \langle x(t), \psi_{ij}(t) \rangle &= \sum \langle x(t), g(k) \frac{1}{2^{(j-1)/2}} \phi(2^{-j+1}(t-2^{-j}i)-k) \rangle \\ &= m_k g(k) \mathcal{A}_{i-k}^{j+1}(x) \end{aligned}$$
(22)



Figure 12. A Daubechies- 8 (D-8) function and its spectral properties: (**a**) D-8 scaling function; (**b**) D-8 wavelet function; (**c**) D-8 scaling Fourier transform; (**d**) D-8 wavelet Fourier transform.

To complete the construction of the filter pair $\{H(\cdot), G(\cdot)\}$, the properties between the approximation subspaces $(\{V_j\}_{j \in \mathbb{Z}})$ and detail subspaces $(\{W_j\}_{j \in \mathbb{Z}})$ are exploited to derive the design criteria for the discrete filters. Specifically, the same scale orthogonality property between the scaling and detail subspaces in

$$\sum_{k\in Z} \Phi(2^{j}\omega + 2k\pi)\Psi(2^{j}\omega + 2k\pi) = 0, \forall j$$
(23)

is the first property that the resulting filters should satisfy. For the sake of illustration, fix j = 1 and use

$$\sqrt{2}\Phi(2\omega) = H(\omega)\Phi(\omega) \tag{24}$$

Using Eq. (24) together with the orthonormality property of *j*th scale basis functions $\{\Phi_{jk}(t)\}$ [4], namely, $\sum |\Phi(\omega + 2k\pi)|^2 = 1$, where we separate even and odd terms, yields the first property of one of the so-called conjugate mirror filters:

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1.$$
 (25)

Using the nonoverlapping property expressed in Eq. (23), together with the evaluations of $\Psi(2\omega)$ and $\Phi(2\omega)$ and making a similar argument to that advanced in the preceding equation yield the second property of conjugate mirror filters:

$$H(\omega)G^{*}(\omega) + H(\omega + \pi)G^{*}(\omega + \pi) = 0.$$
 (26)

The combined structure of the two filters is referred to as "a conjugate mirror filter bank," and their respective impulse responses completely specify the corresponding wavelets (see numerous references, e.g., Ref. 6 for additional technical details). Note that the literature in the MR studies tends to follow one of two possible strategies:

- A more functional analytic approach, which is mostly followed by applied mathematicians/mathematicians and scientists [4,5,16] (we could not possibly do justice to the numerous good texts now available; the author cites only what he is most familiar with.)
- A more filtering-oriented approach widely popular among engineers and some applied mathematicians [7,24,27].

3.4. Refining a Wavelet Basis: Wavelet Packet and Local Cosine Bases

A selected analysis wavelet is not necessarily well adapted to any observed signal. This is particularly the case when the signal is time-varying and has a rich spectral structure.

One approach is to then partition the signal into homogeneous spectral segments and by the same token find it an adapted basis. This is tantamount to further refining the wavelet basis, and resulting in what is referred to as a *wavelet packet basis*. A similar adapted partitioning may be carried out in the time domain by way of an orthogonal local trigonometric basis (sine or cosine). The two formulations are very similar, and the solution to the search for an adapted basis in both cases is resolved in precisely the same way. In the interest of space, we focus in the following development on only wavelet packets.

3.4.1. Wavelet Packets. We maintained above that a selection of an analysis wavelet function should be carried out in function of the signal of interest. This, of course, assumes that we have some prior knowledge about the signal at hand. While plausible in some cases, this assumption is very unlikely in most practical cases. Yet it is highly desirable to select a basis that might still lead to an adapted representation of an apriorily "unknown" signal. Coifman and Meyer [28] proposed to refine the standard wavelet decomposition. Intuitively, they proposed to further partition the spectral region of the details (i.e., wavelet coefficients) in addition to partitioning the coarse and/or scaling portion as ordinarily performed for a wavelet representation. This, as shown in Fig. 13, yields an overcomplete representation of a signal, in other words, a dictionary of bases with a tree structure. The binary nature of the tree as further discussed below, affords a very efficient search for a basis which is best adapted to a given signal in the set.

Formally, we accomplish a partition refinement of, say a subspace U_j by way of a new wavelet basis which includes both its approximation and its details, as shown in Fig. 13. This construction due to Coifman, and Meyer [28] may be simply understood as one's ability to find a basis



Figure 13. Wavelet packet bases structure.

for all the subspaces $\{V_j\}_{j\in\mathbb{Z}}$ and $\{W_j\}_{j\in\mathbb{Z}}$ by iterating the nestedness property. This then amounts to expressing two new functions $\tilde{\psi}_j^0(t)$ and $\tilde{\psi}_j^1(t)$ in terms of $\{\tilde{\psi}_{j-1,k}\}_{(j,k)\in\mathbb{Z}^2}$, an orthonormal basis of a generic subspace U_{j-1} (which in our case may be either V_{j-1} or W_{j-1})

$$\tilde{\psi}_{j}^{0}(t) = \sum_{k=-\infty}^{\infty} h(k) \tilde{\psi}_{j-1}(t-2^{j-1}k)$$
(27)

$$\tilde{\psi}_{j}^{1}(t) = \sum_{k=-\infty}^{\infty} g(k) \tilde{\psi}_{j-1}(t - 2^{j-1}k)$$
(28)

where $h(\cdot)$ and $g(\cdot)$ are the impulse responses of the filters in a corresponding filter bank and the combined family $\{\tilde{\psi}_j^0(t-2^{j}k),\tilde{\psi}_j^1(t-2^{j}k)\}_{(j,k)\in\mathbb{Z}^2}$ is an orthonormal basis of U_j . This, as illustrated in Fig. 13, is graphically represented by a binary tree where, at each of the nodes reside the corresponding wavelet packet coefficients. The implementation of a wavelet packet decomposition is a straightforward extension of that of a wavelet decomposition, and consists of an iteration of both $H(\cdot)$ and $G(\cdot)$ bands (see Fig. 14) to naturally lead to an overcomplete representation. This is reflected on the tree by the fact that, with the exception of the root and bottom nodes, which have only respectively two children nodes and one ancestor node, each node bears two children nodes and one ancestor node.

3.4.2. Basis Search. To identify the best adapted basis in an overcomplete signal representation, as noted above, we first construct a criterion that, when optimized, will reflect desired properties intrinsic to the signal being analyzed. The earliest proposed criterion applied to a wavelet packet basis search is the so-called entropy criterion [29]. Unlike Shannon's information theoretic criterion, this is additive and makes use of the coefficients residing at each node of the tree in lieu of computed probabilities. The presence of more complex features in a signal necessitates such an adapted basis to ultimately achieve an ideally more parsimonious or succinct representation. As pointed out earlier, when searching for a wavelet packet or local cosine best basis, we typically have a dictionary \mathcal{D} of possible bases with a binary tree structure. Each node (j, j') (where $j \in \{0, \dots, J\}$ represents the depth and $j' \in \{0, \ldots, 2^j - 1\}$ represents the



Figure 14. Wavelet packet filter bank realization.

branches on the *j*th level) of the tree then corresponds to a given orthonormal basis $\mathcal{B}_{j,j'}$ of a vector subspace of $\ell^2(\{1,\ldots,N\})$ [$\ell^2(\{1,\ldots,N\})$ is a Hilbert space of finiteenergy sequences]. Since a particular partition $p \in \mathcal{P}$ of [0, 1] is composed of intervals $I_{j,j'} = [2^{-j}j', 2^{-j}(j'+1)]$, an orthonormal basis of $\ell^2(\{1,\ldots,N\})$ is given by $\mathcal{B}^p = \bigcup_{(j,j')|I_{j,j'} \in \mathcal{P}|}\mathcal{B}_{j,j'}$. By taking advantage of the property

$$\operatorname{Span}\{\mathcal{B}_{j,j'}\} = \operatorname{Span}\{\mathcal{B}_{j+1,2j'}\} \stackrel{\scriptscriptstyle{\leftarrow}}{\oplus} \operatorname{Span}\{\mathcal{B}_{j+1,2j'+1}\}$$
(29)

where \oplus denotes a subspace direct sum, we associate to each node a cost $\mathcal{C}(\cdot)$. We can then perform a bottom-up comparison of children versus parent costs (thereby, in effect, eliminating all redundand or inadequate leaves from the tree) and ultimately prune the tree.

Our goal is to then choose the basis that leads to the *best* approximation of $\{x[t]\}$ among a collection of orthonormal bases $\{\mathcal{B}^p = \{\Psi x_i p\}_{1 \le i \le N} | p \in \mathcal{P}\}$, where the term x_i emphasizes that it is adapted to $\{x[t]\}$. Trees of wavelet packet bases studied by Coifman and Wickerhauser [29] are constructed by quadrature mirror filter banks and constitute functions that are well localized in time and frequency. This family of orthonormal bases, partitions the frequency axis into intervals of different sizes, with each set corresponding to a specific wavelet packet basis. Another family of orthonormal bases, studied by Malvar [31], and Coifman and Meyer [28], can be constructed with a tree of windowed cosine functions, and correspond to a division of the time axis into intervals of dyadically varying sizes.

For a discrete signal of size N (e.g., the size of the WP tableau shown in Fig. 13), one can show that a tree of wavelet packet bases or local cosine bases has $P = N(1 + \log_2 N)$ distinct vectors but includes more than $2^{N/2}$ different orthogonal bases. One can also show that the signal expansion in these bases is computed with algorithms that require $O(N \log_2 N)$ operations. Coifman and Wickerhauser [29] proposed that for any signal $\{x[m]\}$ and an appropriate functional $\mathcal{K}(\cdot)$, one finds the best basis \mathcal{B}^{p_0} by minimizing an "additive" cost function

$$\operatorname{Cost}(\mathbf{x}, \mathcal{B}^p) = \sum_{i=1}^N \mathcal{K}(|\langle \mathbf{x}, \Psi x_i p \rangle|^2)$$
(30)

over all bases. As a result, the basis that results from minimizing this cost function corresponds to the "best" representation of the observed signal. The resulting pruned tree of Fig. 15 bears the coefficients at the remaining leaves and nodes.

4. MR APPLICATIONS IN ESTIMATION PROBLEMS

The computational efficiency of a wavelet decomposition together with all its properties have triggered unprecedented interest in their application in the area of information sciences [e.g., 32-37]. Specific applications have ranged from compression [38-41] to signal or image modeling [42-45], and from signal or image enhancement to communications [46-53]. The literature in statistical applications as a whole has seen an explosive growth, and



Figure 15. Tree pruning in search for a "best" basis.

in the interest of space, we will focus our discussion on a somewhat broader perspective that may, in essence, be usefully reinterpreted in a number of different instances.

4.1. Signal Estimation Denoising and Modeling

Denoising may be heuristically interpreted as a quest for parsimony of a representation of a signal. Wavelets as described above, have a great capacity for energy compaction, particularly at or near singularities. Given a particular wavelet, its corresponding basis, as noted above, is not universally optimal for all signals and particularly not for a noisy one; this difficulty may be lifted by adapting the representation to the signal in a "best" way possible and according to some criterion. The first of the two possible ways is to pick an optimal basis in a wavelet packet dictionary and discard the negligible coefficients [54]. The second, which focuses on reconstruction of a signal in noise, and which we discuss here, accounts for the underlying noise statistics in the multiscale domain to separate the "mostly" signal part from the "mostly" noise part [55-57]. We opt here to discuss a more general setting that assumes unknown noise statistics and where a signal reconstruction is still sought in some optimal way, as we elaborate below. This approach is particularly appealing in that it may be reduced to the setting in earlier developments.

4.1.1. Problem Statement. Consider an additive noise model

$$x(t) = s(t) + n(t)$$
 (31)

where s(t) is an unknown but deterministic signal corrupted by a zero-mean noise process n(t), and x(t) is the observed, that is, noisy, signal. The objective is to recover the signal $\{s(t)\}$ based on the observations $\{x(t)\}$.

The underlying signal is modeled with an orthonormal basis representation

$$s(t) = \sum_{i} C_{i}^{s} \psi_{i}(t)$$

and similarly the noise is represented as

$$n(t) = \sum_{i} C_{i}^{n} \psi_{i}(t)$$

By linearity, the observed signal can also be represented in the same fashion, and its coefficients are given by

$$C_i^x = C_i^s + C_i^n.$$

A key assumption we make is that for certain values of $i, C_i^s = 0$; in other words, the corresponding observation coefficients C_i^x represent "pure noise," rather than signal corrupted by noise. As shown by Krim and Pesquet [58], this is a reasonable assumption in view of the spectral and structural differences between the underlying signal s(t)and the noise n(t) across scales. Given this assumption, wavelet-based denoising consists of determining which wavelet coefficients represent primarily signal, and which mostly capture noise. The goal is to then localize and isolate the "mostly signal" coefficients. This may be achieved by defining an information measure as a function of the wavelet coefficients. It identifies the "useful" coefficients as those whose inclusion improves the data explanation. One such measure is Rissanen's information-theoretic approach [or minimum description length (MDL)] [59]. In other words, the MDL criterion is utilized for resolving the tradeoff between model complexity (each retained coefficient increases the number of model parameters) and goodness of fit [each truncated coefficient decreases the fit between the received (i.e., noisy) signal and its reconstruction].

4.1.2. The Coding Length Criterion. Wavelet thresholding is essentially an order estimation problem, one of balancing model accuracy against overfitting, and one of capturing as much of the "signal" as possible, while leaving out as much of the "noise" as possible. One approach to this estimation problem is to account for any prior knowledge available on the signal of interest, which usually is of probabilistic nature. This leads to a Bayesian estimation approach as developed in Refs. 60 and 61. While generally more complex, it does provide a regularization capacity which is much needed in low-SNR environments.

A parsimony-driven strategy, which we expound on here, addresses the problem of modeling in general, and that of compression in particular. It provides in addition a fairly general and interesting framework where the signal is assumed deterministic and unknown, and results in some intuitively sensible and mathematically tractable techniques [54]. Specifically, it brings together Rissanen's work on stochastic complexity and coding length [59,62], and Huber's work on minimax statistical robustness [63,64].

Following Rissasen, we seek the data representation that results in the shortest encoding of both observations and complexity constraints. As a departure from the commonly assumed Gaussian likelihood, we rather assume that the noise distribution f of our observed sequence is a (possibly) scaled version of an unknown member of the family of ε -contaminated normal distributions

$$\mathcal{P}_{\varepsilon} = \{ (1 - \varepsilon)\Phi + \varepsilon G : G \in \mathcal{F} \}$$

where Φ is the standard normal distribution, \mathcal{F} is the set of all suitably smooth distribution functions, and

 $\varepsilon \in (0, 1)$ is the known fraction of contamination. (This is no loss of generality, since ε may always be estimated if unknown.) Note that this study straightforwardly reduces to the additive Gaussian noise case, by setting the mixture parameter $\varepsilon = 0$, and is in that sense more general.

For fixed model order, the expectation of the MDL criterion is the entropy, plus a penalty term that is independent of both the distribution and the functional form of the estimator. In accordance with the minimax principle, we seek the least favorable noise distribution and evaluate the MDL criterion for that distribution. In other words, we solve a minimax problem where the entropy is maximized over all distributions in $\mathcal{P}_{\varepsilon}$, and the description length is minimized over all estimators in \mathcal{S} . The saddle point (provided its existence) yields a minimax robust version of MDL, which we call the minimax description length (MMDL) criterion.

Krim and Schick [54], show that the least favorable distribution in $\mathcal{P}_{\varepsilon}$, which also maximizes the entropy, is one that is Gaussian in the center and Laplacian ("double exponential") in the tails, and switches from one to the other at a point whose value depends on the fraction of contamination ε .

Proposition 1. The distribution $f_H \in \mathcal{P}_{\varepsilon}$ that minimizes the negentropy is

$$f_{H}(c) = \begin{cases} (1-\varepsilon)\phi_{\sigma}(a)e^{(1/\sigma^{2})(ac+a^{2})} & c \leq -a\\ (1-\varepsilon)\phi_{\sigma}(c) & -a \leq c \leq a\\ (1-\varepsilon)\phi_{\sigma}(a)e^{(1/\sigma^{2})(-ac+a^{2})} & a \leq c \end{cases}$$
(32)

where ϕ_{σ} is the normal density with mean zero and variance σ^2 and a is related to ε by the equation

$$2\left(\frac{\phi_{\sigma}(a)}{a/\sigma^2} - \Phi_{\sigma}(-a)\right) = \frac{\varepsilon}{1-\varepsilon}$$
(33)

4.2. Coding for Worst-Case Noise

Let the set of wavelet coefficients obtained from the observed signal be denoted by $C^N = \{C_1^x, C_2^x, \ldots, C_N^x\}$ as a time series without regard to the scale, and where the superscript indicates the corresponding process. Let exactly K of these coefficients contain signal information, while the remainder only contain noise. If necessary, we reindex these coefficients so that

$$C_i^{\alpha} = \begin{cases} C_i^{\alpha} + C_i^{n} & i = 1, 2, \dots, K\\ C_i^{n} & \text{otherwise} \end{cases}$$
(34)

By assumption, the set of noise coefficients $\{C_i^n\}$ is a sample of independent, identically distributed random variates drawn from Huber's distribution f_H . It follows, by Eq. (34), that the observed coefficients C_i^x obey the distribution $f_H(c - C_i^s)$ when i = 1, 2, ..., K, and $f_H(c)$ otherwise. Thus, the likelihood function is given by

$$\ell(\mathcal{C}^N; K) = \prod_{i \le K} f_H(C_i^{\alpha} - C_i^s) \prod_{i > K} f_H(C_i^{\alpha})$$

Since f_H is symmetric and unimodal with a maximum at the origin, this expression is maximized (with respect to the signal coefficient estimates $\{\hat{C}_i^s\}$) by setting

$$\hat{C}_i^s = C_i^x$$

for i = 1, 2, ..., K. It follows that the maximized likelihood (given K) is

$$\ell^*(\mathcal{C}^N;K) = \prod_{i \le K} f_H(0) \prod_{i > K} f_H(C_i^x)$$

Thus, the problem is reduced to choosing the optimal value of *K*, in the sense of minimizing the MDL criterion:

$$\mathcal{L}(\mathcal{C}^{N};K) = -\log \ell^{*}(\mathcal{C}^{N};K) + K\log N$$
$$= -\sum_{i \leq K} \log f_{H}(0) - \sum_{i > K} \log f_{H}(C_{i}^{x}) + K\log N \quad (35)$$

Neglecting terms independent of K, this is equivalent to minimizing

$$ilde{\mathcal{L}}(\mathcal{C}^N;K) = rac{1}{2\sigma^2}\sum_{i>K}\eta(C^{\mathrm{x}}_i) + K\log N$$

where

$$\eta(c) = \begin{cases} c^2 & \text{if } |c| < a \\ a|c| - a^2 & \text{otherwise} \end{cases}$$

is proportional to the exponent in Huber's distribution f_H . This can simply be achieved by a thresholding scheme [54].

Proposition 2. When $\log N > a^2/2\sigma^2$, the coefficient $|C_i^{x}|$ is truncated if

$$\begin{split} |C_i^{\mathrm{x}}| &< \frac{a}{2} + \frac{\sigma^2}{a} \log N \qquad (case \ 1) \end{split}$$

When $\log N \leq \frac{a^2}{2\sigma^2}$, the coefficient $|C_i^{\mathrm{x}}|$ is truncated if
 $|C_i^{\mathrm{x}}| < \sigma \sqrt{2\log N} \qquad (case \ 2)$

Remarks. More ample details may be found in the paper by Krim and Schick [54]. Note however, when $\sigma^2 \rightarrow 0$, the thresholding scheme reduces to case 2, and C_i^x is never truncated; since this represents the no-noise case, it is reasonable that all coefficients should be retained in the reconstruction. On the other hand, for large σ^2 , the thresholding scheme reduces to case 1, which is more conservative. For $\sigma^2 \rightarrow \infty$, the signal-to-noise ratio becomes zero and the best one can do is to estimate the signal as identically zero.

Similarly, when $a \to \infty$, the noise distribution becomes purely Gaussian, and the thresholding scheme reduces to case 2, as expected. The resulting threshold of this particular noise case coincides with the results of Donoho and others [55,56] and is qualitatively similar to that derived by Saito [57]. On the other hand, when $a \to 0$, the noise distribution becomes purely Laplacian, and the thresholding scheme reduces to Case 1. Finally, when $N \rightarrow 1$, the thresholding scheme reduces to case 2, suggesting that outliers are unlikely to occur in a small sample, and it is hence more reasonable to assume purely Gaussian noise. On the other hand, for large N, the thresholding scheme reduces to case 1, since outliers are highly likely to occur in a large sample.

It is important to distinguish the minimax error result obtained by Donoho and Johnstone [55], which was achieved over a signal smoothness class, from those discussed here and derived by Krim and Schick [54], which are obtained over a family of noise distributions.

4.2.1. Numerical Experiments. In the examples that follow, we demonstrate the performance of the robust thresholding procedure described above, and compare it with that of the thresholding scheme based on the assumption of normally distributed noise.

Example 1. Using WAVELAB (available from the Stanford Statistics Department, courtesy of D. L. Donoho and I. M. Johnstone), we synthesized a broken ramp signal of length N = 1024. This signal admits an efficient representation in a wavelet basis, that is, one with very few nonzero coefficients. The noise is additive and independent identically distributed, obeying a $N(0, \sigma^2)$ distribution contaminated by a fraction $\varepsilon = 10\%$ of white Gaussian noise with distribution $N(0, 9\sigma^2)$. The overall signal-to-noise ratio (SNR) was maintained at 10 dB (see Fig. 16, top).

We implemented two estimators, the first of which is based on a purely Gaussian noise assumption (i.e., $\varepsilon = 0$), and where the thresholding scheme due to [58,65] was used. The second was the MMDL robust estimator described above. The reconstructions based on each estimator appear in Fig. 16. As may easily be observed, and in contrast to the MMDL technique, the Gaussian assumption induces a high susceptibility to outliers.

Monte Carlo simulations were carried out to evaluate the reconstruction performance over a range of SNRs. At each value of the SNR, 100 experiments were conducted, and the cumulative reconstruction error is displayed in Fig. 17. The robust estimator uniformly outperforms the *classic* estimator in both L_1 and L_2 errors over a wide range of SNRs. Furthermore, the performances of the Gaussian and robust estimators become indistinguishable at high SNRs, that is, with small noise variance, showing that robustness does not come at the cost of reduced efficiency.

Bounding the Reconstruction Error. Although the robust estimator-based reconstruction error is much improved it is still potentially unbounded. As discussed further by Krim and Schick, [54], and because of the compactness of wavelets, unbounded noise will still result in unbounded reconstruction error, a property that may be considered undesirable. This problem may be circumvented by making the assumption that the signal has bounded energy; in that case, one of at least two alternatives is possible:

1. In practice, the signal is known to be bounded, and prior knowledge of the physical properties of the signal may be used to determine the $\|\cdot\|_{\infty}$ of the sequence of signal wavelet coefficients $\{C_i^s\}$.



Figure 16. Noisy ramp signal and its Gaussian and robust reconstructions.



Figure 17. L_1 and L_2 error performance versus SNR, for the Gaussian and robust estimators.

This information may be used to truncate observed coefficients $\{C_i^x\}$ not only below, as discussed earlier, but also above.

2. In the absence of such prior knowledge, it may still be possible to bound the reconstruction error through



Figure 18. Absolute reconstruction error versus outlier amplitude for three thresholding schemes.

an adaptive supremum-secondary thresholding scheme based on some representation criterion, such as entropy.

The first of these approaches is illustrated in Fig. 18, which uses the following modified thresholding. Let $\alpha > 0$ be an upper bound on the magnitude of the signal coefficients; then

$$ilde{C}^{\mathrm{s}}_i = egin{cases} 0 & ext{if } |C^{\mathrm{s}}_i| \leq rac{a}{2} + rac{\sigma^2}{a} \log N \ \hat{C}^{\mathrm{s}}_i = C^{\mathrm{s}}_i & ext{if } rac{a}{2} + rac{\sigma^2}{a} \log K \leq |C^{\mathrm{s}}_i| \leq lpha \ lpha \operatorname{sgn}(C^{\mathrm{s}}_i) & ext{if } lpha \leq |C^{\mathrm{s}}_i| \end{cases}$$

provided $\log N > a^2/2\sigma^2$ and $\alpha > (a/2) + (\sigma^2/a) \log N$. The graph shows that although the robust estimator's reconstruction error initially grows more slowly than that of the Gaussian estimator, the two errors soon converge as the variance of the outliers grows. The reconstruction error for the bounded-error estimator, however, levels off past a certain magnitude of the outlier, as expected.

Sensitivity Analysis for the Fraction of Contamination. Although such crucial assumptions as the normality of the noise or exact knowledge of its variance σ^2 usually go unremarked, it is often thought that Huber-like approaches are limited on account of the assumption of known ε . We demonstrate the resilience and robustness of the approach by studying the sensitivity of the estimator to changes in the assumed value of ε .

Figure 19 shows the total reconstruction error as a function of variation in the true fraction of contamination ε . In other words, an abscissa of 0 corresponds to an assumed fraction of contamination equal to the true fraction; larger abscissas correspond to outliers of larger magnitude than assumed by the robust estimator, and vice versa. Clearly, the Gaussian estimator assumes zero contamination throughout. Figure 19 shows that the reconstruction error for the Gaussian estimator grows very rapidly as the true fraction of contamination increases, whereas that of the robust estimator is nearly flat over a broad range. This should not come as a surprise: outliers



Figure 19. Error stability versus variation in the mixture parameter ε .

are, by definition, rare events, and for a localized procedure such as wavelet expansion, the precise frequency of outliers is much less important than their existence at all.

Example 2. Example 1 assumed a fixed wavelet basis. As discussed above, however, highly nonstationary signals can be represented most efficiently by using an adaptive basis that automatically chooses resolutions as the behavior of the signal varies over time. Using a L^2 error criterion, we search for the best basis of a chirp and show the reconstruction in Fig. 20 (see Krim et al. [66] for more details).

5. CONCLUSION

1000

500

-500

-1000

-1500

s[*m*]

We have given an overview of an already broad area of research and discussed its application in information sciences, and more specifically an estimation technique that we believe captures the essence of many encountered problems. The article assumes an advanced undergraduate knowledge in electrical engineering applications and mathematics and may also play a role of a primer for a first-time reader on the topic and is aimed at providing a working knowledge of the tools, deferring most of the technical details to the references. While the bibliography is certainly incomplete (too large for it to be exhaustive), it hopefully provides a sufficient overview for the interested reader who may want to probe further.

BIOGRAPHY



Hamid Krim received his degrees in electrical engineering from University of Washington and Northeastern University. In 1991 he became a NSF postdoctoral scholar

Figure 20. A noisy chirp reconstructed from its best basis and denoised as shown in the figure on the right.

at Foreign Centers of Excellence (LSS Supelec/Univ. of Orsay, Paris, France). He subsequently joined the Laboratory for Information and Decision Systems, MIT, Cambridge, Massachusetts, as a research scientist performing/supervising research in his area of interest and was an original contributor to the Center for Imaging Science sponsored by ARO. He then joined the faculty in the ECE Department at North Carolina State University in Raleigh, North Carolina in 1998. He also is a recipient of the NSF Career Young Investigator Award and an associate editor of the IEEE Trans. On SP. As a member of technical staff at AT&T Bell Labs, he has worked in the area of telephony and digital communication systems/subsystems. His research interests are in statistical estimation and detection and mathematical modeling with a keen emphasis on applications.

BIBLIOGRAPHY

- 1. E. Brigham, *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
- 2. A. Papoulis, Signal Analysis, McGraw-Hill, New York, 1977.
- D. Gabor, Theory of communication, J. IEE 93: 429-457 (1946).
- 4. S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, Boston, 1997.
- 5. I. Daubechies, S. Mallat, and A. Willsky, Special edition on wavelets and applications, *IEEE Trans. on IT*, 1992.
- I. Daubechies, Ten Lectures on Wavelets, CBMS-NSF, SIAM, Philadelphia, 1992.
- 7. M. Vetterli and J. Kovacevic, Wavelets and Subband Coding, Prentice-Hall, Englewood Cliffs, NJ, 1995.
- 8. H. Krim, On the distribution of optimized multiscale representations, in *ICASSP*, Vol. V, (Munich, Germany), IEEE, May 1997.
- 9. G. H. Golub and C. F. VanLoan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1984.
- D. Luenberger, Optimization by Vector Space Methods, Wiley, New York, 1968.
- K. Gröchenig, Acceleration of the frame algorithm, *IEEE Trans. Signal Process.* 41: 3331-3340 (Dec. 1993).
- A. B. Hamza and H. Krim, Image Denoising: A Nonlinear Robust Statistical Approach, *IEEE*, *Trans. SP* **SP-49**(12): 3045-3054 (2001).
- Y. Meyer, Ondelettes et opérateur, Vol. 1, Hermann, Paris, 1990.
- S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Trans. Pattern Anal. Mach. Int.* PAMI-11: 674–693 (July 1989).
- Y. Meyer, Wavelets and Applications, SIAM, Philadelphia, 1992.
- Y. Meyer, Ondelettes et opérateur, Vol. 1, Hermann, Paris, 1990.
- F. J. Hampson and J.-C. Pesquet, A Nonlinear Decomposition with Perfect Reconstruction, IEEE, Trans. on IP 7-11, 1998, 1547–1560.
- P. L. Combettes and J.-C. Pesquet, Convex multiresolution analysis, *IEEE Trans. Pattern Anal. Mach. Int.* 20: 1308–1318 (Dec. 1998).

 W. Sweldens, The lifting scheme: A construction of second generation wavelets, SIAM J. Math. Anal. 29(9): 511–546 (1997).

2861

- S. Mallat, Multiresolution approximation and wavelet orthonormal bases of L²(ℝ), Trans. Am. Math. Soc. 315: 69-87 (Sept. 1989).
- J. Woods and S. O'Neil, Sub-band coding of images, *IEEE Trans. ASSP* 34: 1278-1288 (May 1986).
- M. Vetterli, Multi-dimensional subband coding: some theory and algorithms, Signal Process. 6: 97–112 (April 1984).
- F. Meyer and R. Coifman, Brushlets: a tool for directional image analysis and image compression, *Appl. Comput. Harm. Anal.* 147–187 (1997).
- 24. G. Strang and T. Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge Press, Boston, 1996.
- P. P. Vaidyanathan, Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial, *Proc. IEEE* 78: 56–93 (Jan. 1990).
- A. Cohen, I. Daubechies, and J. C. Feauveau, Biorthogonal bases of compactly supported wavelets, 1992. Unpublished Technical Report.
- P. P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice-Hall, Englewood Cliffs, NJ, 1992.
- R. Coifman and Y. Meyer, Remarques sur l'analyse de Fourier à fenêtre, C. R. Acad. Sci. Série I 259-261 (1991).
- R. R. Coifman and M. V. Wickerhauser, Entropy-based algorithms for best basis selection, *IEEE Trans. Inform. Theory* **IT-38**: 713-718 (March 1992).
- M. V. Wickerhauser, INRIA lectures on wavelet packet algorithms, in On-delettes et paquets d'ondelettes INRIA T-R (Roquencourt), June 17-21, 1991, pp. 31-99.
- H. Malvar, Lapped transforms for efficient transform subband coding, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-38: 969-978 (June 1990).
- A. Aldroubi and E. M. Unser, Wavelets in Medicine and Biology, CRC Press, Boca Raton, FL, 1996.
- A. Akansu and E. M. J. Smith, Subband and Wavelet Transforms, Kluwer, 1995.
- 34. A. Arneodo, F. Argoul, J. E. E. Bacry, and J. Muzy, Ondelettes, Multifractales et Turbulence, Diderot, Paris, 1995.
- A. Antoniadis and E. G. Oppenheim, Wavelets and Statistics, Vol. LNS 103, Lecture Notes in Statistics, Springer-Verlag, 1995.
- P. Mueller and E. B. Vidakovic, Bayesian Inference in Wavelet-Based Models, Vol. LNS 141, Lecture Notes in Statistics, Springer-Verlag, 1999.
- B. Vidakovic, Statistical Modeling by Wavelets, Wiley, New York, 1999.
- K. Ramchandran and M. Vetterli, Best wavelet packet bases in a rate-distorsion sense, *IEEE Trans. Image Process.* 2: 160-175 (April 1993).
- J. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. Signal Process.* 41: 3445–3462 (1993).
- P. Cosman, R. Gray, and M. Vetterli, Vector quantization of image subbands: a survey, *IEEE Trans. Image Process.* 5: 202-225 (Feb. 1996).

2862 WDM METROPOLITAN-AREA OPTICAL NETWORKS

- A. Kim and H. Krim, Hierarchical stochastic modeling of sar imagery for segmentation/compression, *IEEE Trans. SP* 458-468 (Feb. 1999).
- M. Basseville et al., Modeling and estimation of multiresolution stochastic processes, *IEEE Trans. Inform. Theory* **IT-38**: 529–532 (March 1992).
- M. Luettgen, W. Karl, and A. Willsky, Likelihood calculation for multiscale image models with applications in texture discrimination, *IEEE Trans. Image Process.* 3(1): 41-64 (1994).
- E. Fabre, New fast smoothers for multiscale systems, *IEEE Trans. Signal Process.* 44: 1893–1911 (Aug. 1996).
- P. Fieguth, W. Karl, A. Willsky, and C. Wunsch, Multiresolution optimal interpolation and statistical ananlysis of topex/poseidon satellite altimetry, *IEEE Trans. Geosci. Remote Sens.* 33: 280-292 (March 1995).
- 46. M. K. Tsatsanis and G. B. Giannakis, Principal component filter banks for optimal multiresolution analysis, *IEEE Trans. Signal Process.* 43: 1766–1777 (Aug. 1995).
- 47. M. K. Tsatsanis and G. B. Giannakis, Optimal linear receivers for ds-cdma systems: A signal processing approach, *IEEE Trans. Signal Process.* 44: 3044–3055 (Dec. 1996).
- A. Scaglione, G. B. Giannakis, and S. Barbarossa, Redundant filterbank precoders and equalizers, parts i and ii, *IEEE Trans. Signal Process.* 47: 1988–2022 (July 1999).
- A. Scaglione, S. Barbarossa, and G. B. Giannakis, Filterbank transceivers optimizing information rate in block transmissions over dispersive channels, *IEEE Trans. Inform. Theory* 45: 1019-1032 (April 1999).
- R. Learned, H. Krim, A. Willsky, and W. Karl, Waveletpacket-based multiple access communication, *Wavelet Appl.* Signal Image Proc. II: 246-259 (Oct. 1994).
- R. Learned, A. Willsky, and D. Boroson, Low complexity optimal joint detection for oversaturated multiple access communications, *IEEE Trans. Signal Process.* 45(1): 113-123 (1997).
- A. Lindsey, Wavelet packet modulation for orthogonally multiplexed communication, *IEEE Trans. Signal Process.* 45(5): 520-524 (1997).
- 53. K. Wong, J. Wu, T. Davidson, and Q. Jin, Wavelet packet division multiplexing and wavelet packet design under timing error effects, *IEEE Trans. Signal Process.* 45: 2877–2886 (Dec. 1997).
- 54. H. Krim and I. Schick, Minimax description length for signal denoising and optimized representation, *IEEE Trans. Inform. Theory* (April 1999). (Eds. H. Krim, W. Willinger, A. Iouditski and D. Tse).
- D. L. Donoho and I. M. Johnstone, *Ideal Spatial Adaptation* by Wavelet Shrinkage, preprint, Dept. Statistics, Stanford Univ., June 1992.
- H. Krim, S. Mallat, D. Donoho, and A. Willsky, Best basis algorithm for signal enhancement, in *ICASSP'95*, Detroit, MI, IEEE, May 1995.
- N. Saito, Local Feature Extraction and Its Applications Using a Library of Bases, Ph.D. thesis, Yale Univ., Dec. 1994.
- H. Krim and J.-C. Pesquet, On the statistics of best bases criteria, in Wavelets in Statistics, Lecture Notes in Statistics, Springer-Verlag, July 1995.
- J. Rissanen, Modeling by shortest data description, Automatica 14: 465–471 (1978).

- B. Vidakovic, Nonlinear wavelet shrinkage with bayes rules and bayes, J. Am. Stat. Assoc. 93: 173–179 (1998).
- D. Leporini, J.-C. Pesquet, and H. Krim, Best Basis Representation with Prior Statistical Models, Lecture Notes in Statistics, LNS 141 Springer-Verlag, 1999, Chap. 11.
- J. Rissanen, Stochastic complexity and modeling, Ann. Stat. 14: 1080–1100 (1986).
- P. Huber, Robust estimation of a location parameter, Ann. Math. Stat. 35: 1753-1758 (1964).
- 64. P. Huber, *Théorie de l'inférence statistique robuste*, technical report, Univ. Montreal, Montreal, Quebec, Canada, 1969.
- D. Donoho and I. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, J. Am. Stat. Assoc. 90: 1200–1223 (Dec. 1995).
- H. Krim, D. Tucker, S. Mallat, and D. Donoho, Near-optimal risk for best basis search, *IEEE Trans. Inform. Theory* 45(7): 2225–2238 (Nov. 1999).

WDM METROPOLITAN-AREA OPTICAL NETWORKS

IOANNIS TOMKOS Athens Information Technology Peania, Greece

1. INTRODUCTION TO WDM OPTICAL NETWORKS

The role of a telecommunications network is to transmit information in the most efficient, reliable, and costeffective way. To do this, a communications channel with enough bandwidth to satisfy the traffic demands is needed. From all the known transmission media, the optical fiber has the largest available bandwidth (>50 THz) and can satisfy in the most efficient way the traffic demands. Therefore, fiberoptic links are the backbone of current high-speed communication networks.

To take full potential of the available fiber bandwidth, several multiplexing techniques have been introduced in optical communication systems in accordance with conventional digital communications systems. Among all of them, the most popular is the wavelength-division multiplexing (WDM) technique [1]. WDM allows multiple datastreams from various application protocols to be combined and transmitted in parallel at different wavelengths, thereby multiplying the capacity of a single fiber strand. The different wavelengths, separated from one another at a fixed spacing frequency, are multiplexed together using a WDM multiplexer and are then transmitted over the same optical fiber. The signals at the output end of the fiber are demultiplexed and are redistributed into the various applications. Depending on the application, WDM can be deployed in a coarse version (CWDM) with 16 or fewer wavelengths having relatively wide spacing or in a dense version (DWDM) with up to hundreds of wavelengths. The state-of-the-art commercially available DWDM systems utilize 80 10-Gbps (gigabit per second) channels spaced at 50-GHz-frequency separation. The next generation of DWDM systems will eventually utilize 160 channels spaced at 25 GHz.

The first commercial systems that used fiber to transmit signals instead of WDM utilized space divisionmultiplexing (SDM) through the use of multiple fibers and single channel transmission per fiber (Fig. 1a). To overcome losses, optoelectronic (OEO) regenerators were frequently used, resulting in huge system costs. Several installed systems based on *synchronous optical network/synchronous digital hierarchy* (SONET/SDH) are still using such a technique. The breakthrough for the optical transmission technology came with the use of WDM and the invention of optical amplifiers. Tens of optical channels can be amplified simultaneously, leading to an enormous reduction of cost, by eliminating many fibers and more importantly OEO regenerators (Fig. 1b).

Since the introduction of point-to-point WDM systems, the potential for WDM all-optical networking, without the use of optoelectronic conversions, has been exploited by many research groups [2–6], and is now becoming a commercial reality. Beyond a higher capacity in a single fiber, WDM optical networks offer the capability for transparent wavelength routing [5,6]. Their key network elements are optical switching devices (optical add/drop



Figure 1. Layout of (a) early systems that used fiber to transmit signals, and (b) WDM point-to-point systems with optical amplification.

2864 WDM METROPOLITAN-AREA OPTICAL NETWORKS

multiplexers and optical cross-connects) that enable routing at the wavelength level without the use of OEO conversion [7].

Optical add/drop multiplexers (OADMs) are used to add/remove signals from a WDM "comb" transmitted along a fiber connection. (Fig. 2a). The OADM architectures should introduce minimum impairment to the passthrough signals while enabling access to all the channels in the transmitted fiber with minimum cost [8]. The use of OADMs in transparent chains [9] or in single transparent rings [10] depending on the application, has been discussed. Optical cross-connects (OXCs) are under development and in the future will enable transparent ring interconnection and mesh optical networking [5-7]. OXCs are a more generalized form of OADMs, and besides their use for add/drop of individual signals from a WDM comb, they are used to cross-connect signals coming from different fibers (Fig. 2b). The main functions of the OXC will be to dynamically reconfigure the network—at the wavelength level - for restoration purposes, or to accommodate changes in bandwidth demand.



Figure 2. Schematic representation of the functionality of an (a) OADM and (b) OXC.

In the near future, *wavelength-routed* WDM optical networks will span across all network segments. In a general end-to-end connectivity picture, the current networks consists of three major segments, the longhaul network, the metropolitan-area network, and the residential access network. Figure 3 presents a schematic representation of the layout of the various network segments. A metropolitan-area network (MAN) is simply defined as the part of the network that interfaces between the end users ("residential access" or "last-mile" networks) and the backbone long-haul network [10].

Significant network growth has been observed in metropolitan areas, mainly due to the increased growth in data and IP (Internet Protocol) traffic demands enabled by the introduction of broadband access technologies $(e.g.,\ cable\ modems,\ ADSL/VDSL)$ to end users. Driven mainly by the increasing traffic demand in metropolitan areas [11], WDM is now beginning to expand from a network core technology toward the metropolitan and access network arenas. The focus of research has shifted toward WDM metropolitan networks [10-17], with the goal of bringing the benefits (cost and network efficiency) of the optical networking revolution toward the end users. From both technical and economic perspectives, the ability to provide potentially unlimited transmission capacity is the most obvious advantage of DWDM technology. However, the use of WDM as simply a network infrastructure tool that provides increased system capacity is not as compelling for short-haul networks, and must bring more benefits to the operator to gain acceptance. Bandwidth aside, the most compelling technical advantages of WDM networking for metro (MAN) applications can be summarized as follows:

• *Transparency*. A real catalyst behind the use of WDM in short-haul networks may be its promise of "transparency" in offering new high-end wavelength-based services. Several "shades" of transparency have been envisioned, spanning the spectrum from "full" transparency (format, protocol, bit rate) to some subset. WDM can transparently support at their native bit rate, both time-division multiplexing (TDM) and data formats such as *asynchronous transfer mode* (ATM), *Gigabit Ethernet* (GbE), *Fiber Distributed Data Interface* (FDDI), and



Figure 3. Schematic representation of the layout of the various networks segments.

Enterprise System Connectivity (ESCON). Networks would potentially have the flexibility to transport any kind of data without regard for the restrictions of the SONET digital hierarchy. Full (or analog) transparency at this level of the network would "futureproof" the infrastructure against bit rate increase and new traffic types, and would reduce the equipment load in the signal path, resulting in significant cost advantage.

- *Scalability*. DWDM can leverage the abundance of dark fiber in many metropolitan-area networks to quickly meet demand of capacity on point-to-point links and on spans of existing SONET/SDH rings.
- Dynamic Provisioning. Fast, simple and dynamic provision of network connections gives providers the ability to provide high-bandwidth services to enterprises/end-users in days rather than months.
- *Bandwidth Allocation*. Ability to drop individual wavelengths to the building and utilization of the available bandwidth more efficiently.

In the following text we review the requirements, architectures, and performance issues related with WDM metropolitan area optical networks. We present our considerations about the evolution of MANs. We also discuss the optical impairments that limit the transparency in metropolitan WDM networks and present the characteristics of MAN-optimized optical amplifiers, fibers, and architectures of optical add/drop multiplexers that enable flexible and highly performing metropolitanarea networks.

2. METROPOLITAN OPTICAL NETWORKS: CHARACTERISTICS, DEFINITIONS, AND REQUIREMENTS

The main role of the metropolitan-area network segment is to provide traffic grooming and aggregation of a full range of client protocols from enterprise/private customers in access networks to backbone service provider networks. In addition, since the majority of the traffic stays within the same area, metro networks need to provide efficient networking capabilities within the metropolitan-area.

The design of metropolitan-area networks presents many engineering challenges, especially in light of the large existing base of legacy SONET/SDH infrastructure prevalent in current MANs. These traditional TDM networks were originally designed to transport a limited set of traffic types, mainly multiplexed voice and private line services. SONET/SDH systems are not able to transport efficiently data-optimized protocols (e.g., to carry a GbE signal at 1.25 Gbps, a full 2.488-Gbps SONET/SDH channel is needed-a waste of bandwidth). However, today's MAN market is being driven by the need to streamline network efficiencies under rapidly growing capacity demands and increasingly variable traffic patterns. Hence there is a strong desire to migrate from the current SONET/SDH-based network architecture into a more proactive (dynamic and intelligent), multiservice optical network. This will allow service providers to circumvent the need to perform forklift upgrades or lay more fiber (which is time- and cost-intensive), thereby cost-effectively migrating toward a "futureproof" network.

From an architecture point of view, the conventional metropolitan networks can be separated in two different segments with distinct roles: the "metro-access" (or "collector"), and "metro-core" [or "interoffice" (IOF)] networks (Fig. 4). "Metro-access" networks are responsible for collecting the traffic and forward it to a hub node of the "inter office" network, which in turn will act to network the traffic between hub nodes and redirect it to the backbone long-haul network. The typical length of longest path in IOF networks is \sim 300 km and \sim 100 km in "metro-access" networks. IOF networks are designed mostly as physical rings with a meshed traffic pattern as shown in Fig. 4b. The dots on the schematic indicate network node sites, which perform aggregation of highbandwidth optical connections, cross-connect, and handoff to the core network. Cross-connects at these nodes are quite likely to employ electronic switch cores to support subwavelength "grooming" of circuits. Grooming allows for the most efficient utilization of network bandwidth by performing a mix of space switching (port-to-port) and time-domain switching (time slot to time slot), which cannot be implemented easily in the optical domain. Metro-access networks are designed mostly as physical rings with a hubbed traffic pattern as shown in Fig. 4c. The hub node has access to all traffic present in the network, representing an aggregation point and a connection to the IOF network. "Edge boxes" sitting at the access network edge, perform traffic aggregation and WDM multiplexing of low-bandwidth services.

Metropolitan networks are subject to specific requirements and traffic demands that are quite different from those of the backbone long-haul network. Therefore, metro network carriers face additional challenges due to the distinct characteristics of MANs:

- They are very *cost-sensitive*, since the overall network cost is divided into a smaller number of customers than in the long-haul network.
- They are very *sensitive to space and power consumption* characteristics of the network equipment, since the network carriers are struggling to reduce the cost of operating and maintaining the network.
- They are characterized by more *rapidly changing traffic patterns* requiring fast provisioning and therefore ability for network scalability, modularity, fast reconfiguration, and availability of capacity.
- Since the metro network interfaces with a wide variety of end customers, it needs to support very *diverse types of traffic* (SONET/SDH, Ethernet, ESCON, Fiber Channel, ATM, IP, etc.). Therefore network nodes especially at the edges of the metro network should perform traffic aggregation.
- The bit rates of the data tributaries accessing the "metro-access" network can also vary quite significantly (e.g., from OC-3 to OC-196 for SONET and from 100 Mbps to 10 Gbps for Ethernet traffic). Therefore network nodes should also perform traffic grooming to improve the efficiency of the transport



Figure 4. Metro network reference architecture (a) and traffic pattern for the IOF (b) and metro-access (c) network segments

network by combining the low-bit-rate signals to a high-bit-rate wavelength channel. Moreover, the network should be able to support bandwidth provisioning with various levels of granularity.

• Metropolitan-area optical networks should transport information while satisfying the service-layer agreements for *quality of service and high signal integrity*.

3. SYSTEM OFFERINGS FOR METRO NETWORKS

On the basis of the abovementioned requirements, we can conclude that the system offerings for metro networks should

- Utilize cost-effective devices with small size and low power consumption. The devices may trade these characteristics for lower performance, since the size of the metro networks is relatively small and high signal quality can be preserved.
- Provide many different protocol interfaces and the capability of traffic grooming and aggregation.

- Utilize technologies (e.g., optical switching) that will enable the network carriers to move from costly timeconsuming static network provisioning (the stepby-step process of making an optical connection between two sites) to fast automated provisioning that requires minimum cost. With the proper network management system, the operator should "see" the network resources from a single screen, change the connections through software tools, and receive instant verification that the connection is intact (*point-and-click provisioning*).
- Provide network protection and restoration as major requirements that ensure quality of service [18]. Next-generation optical networks will eventually need to support optical protection switching capabilities, which have proved to considerably reduce the network costs as opposed to protection switching in the electronic domain [19]. The simplest protection option is called "dedicated" or "1 + 1," where two copies of the optical signals are counterpropagated through the network on different fibers and both are delivered to the customer equipment, where the signal having the best quality is detected [20].

More advanced protection schemes perform reuse of capacity (i.e., wavelengths) [19]. For the implementation of optical protection there are several enabling technologies such as fast and reliable optical switches, signal-quality monitors, and transient gain-controlled optical amplifiers [19–20].

Several companies are currently researching and developing systems optimized for use in metropolitan applications. The system offerings that are considered for deployment in metro networks can be separated in broad categories that are listed below:

1. Legacy SONET/SDH. This is the dominant method of optical transport in metro networks today. Roughly 80,000 metro SONET rings exist only in North America today. It is a highly reliable ring-based solution, which offers service protection within 50 ms in case of network failure. It was optimized for voice traffic and is therefore inflexible and inefficient with data traffic. Network connectivity is established with SONET add/drop multiplexers (ADMs) and digital electronic cross-connects (DXCs) for interconnecting the rings (Fig. 4a). The SONET ADMs and DXCs require optical-electrical-optical (O-E-O) conversion, which makes the technology costly. The increasing traffic demands in metropolitan areas have been satisfied by increasing the SONET channel bit rate (i.e., from OC-3 to OC-196) or the number of fibers connecting the nodes, introducing in that case network overlays.

2. *Next-Generation SONET/SDH*. This is the response of SONET/SDH system integrators to the evolving characteristics of metro networks. These are "data-optimized" SONET/SDH systems that are fully interoperable with legacy SONET/SDH networks. They are very compelling solutions since they simplify the network provisioning and improve network efficiency by integrating traffic grooming and aggregation equipment.

3. Metro-Core Point-to-Point DWDM. These systems transport multiple streams (32-128 wavelengths) of optical signals on a single fiber. These conventional WDM solutions have simply translated the SONET ring model to a WDM platform with multiple wavelengths per fiber and possibly optical amplifiers to overcome losses (Fig. 5a). They are typically used to "connect" large carrier offices within the metropolitan area in a point-to-point fashion. The network savings arising from these system offerings can be translated in a first approximation to a reduction of the number of required optical fibers and SONET network overlays that are needed to accommodate the traffic demands. The additional cost of WDM system offerings are in the use of multiplexers/demultiplexers, amplifiers, and devices for signal conditioning management (e.g., variable optical attenuators, dispersion compensating modules).

4. *Metro Core Ring DWDM*. These systems utilize optical add/drop multiplexers, which replace SONET add/drop multiplexers to build physical rings carrying WDM signals (Fig. 5b). OADMs, through the wavelength routing concept, can transform a physical ring topology into any type of network logical topology (ring, mesh, or star). These systems are capable to transport multiple streams (e.g., 32–128 wavelengths) of optical signals on a single fiber. Some systems also promise the use of OXC for transparent ring interconnection (Fig. 5c). In such case, no time-division multiplexing/demultiplexing will be performed at the XCs. This will result in inability to perform grooming, as opposed to DXC, and may delay their introduction in metro networks.

5. Metro-Access WDM. These systems transport optical signals between large enterprise sites and carrier central offices. Shorter distance (typically 5-20 km) and lower capacity requirements mean that these systems need lower optical performance than do metro-core DWDM systems. Coarse WDM (CWDM) systems with channel spacing as high as 20 nm is a suitable technology for such application. Efficient aggregation and transport of data traffic is again a key objective.

6. Next-Generation Data Transport Systems. These are data-only devices (layer 2 switches and layer 3 routers) that will eliminate SONET/SDH systems altogether. They achieve more efficient transport of data traffic, at the expense of limited voice capabilities and questionable reliability. They are based mostly on *Ethernet*, although packet-over-SONET and other transport protocols have been proposed [21]. Ethernet is a transport solution that is almost 100% predominant within the local-area network (LAN) market. Its widespread extension into the metro area is considered to be quite reasonable in the near future. New technologies, such as multiprotocol label switching (MPLS) and resilient packet rings (RPRs), which have been introduced to deliver efficient routing and improved quality of service, are showing significant promise. RPR is a *medium-access control* (MAC) protocol designed to optimize bandwidth utilization and facilitate services over a ring network (unlike Ethernet), while providing carrier-class attributes such as 50-ms ringprotection switching. It is worth pointing out that WDM systems are interoperable with data transport systems and can be used for network upgrades. For example, the network operator can keep TDM voice services over SONET on one wavelength, while deploying the dataoriented technology on another wavelength.

7. All-Optical Packet Switching. This is a dataoptimized transport technology that promises to provide high throughput, packet-level switching, rich routing functionality, and excellent flexibility, making this system offering ideal for metro networks [22–25]. It will utilize OADMs and OXCs that, on top of wavelength switching, will be able to perform packet-level switching. Although this type of system has been researched extensively in the past by many groups and consortia [22–25], it remains the most forward-looking solution.

4. ENGINEERING OF TRANSPARENT METRO NETWORKS

Clearly, optical switching through the use of OADMs and OXCs is a promising technology for provisioning, protection, and restoration of high-bandwidth services in the optical layer. However, the adoption of optical switching technology and optically transparent network architectures will not be automatic by all network carriers. The reduction of O-E-O interfaces presents a special set





WDM optical Add/Drop multiplexer (OADM)

of challenges to network designers. The transparency that WDM systems offer limits the scalability (in terms of the number of channels, bit rate, etc.) or geographic extent of the network. The limitations arise from

• Accumulation of Transmission and Networking Impairments. Several effects, unique in optical transmission and networking systems, limit the maximum physical size of the network that can be supported [10,14,15,26-32]. In a transparent optical network the WDM signals pass through many optical components that may introduce deterioration of signal quality.

• Difficulty in Engineering the Network. Indeed, the network would have to be engineered for the "worst case" [15,26]. As a consequence, the worst path (e.g., the longer restoration path) will have to be known from the outset, and all components must be specified for this worst path (and the specifications will be tighter, the greater the network reach). Furthermore,
transparency requires that the whole network be engineered at once. And once engineered, the network cannot be extended beyond its intended design.

• Difficulty in Performance Monitoring. Performance monitoring of the WDM signals at the bit level is difficult without optoelectronic conversions [33,34]. Also, in-band management information is difficult to add and monitor, which makes it difficult to manage networks with transparent all-optical nodes.

A significant consideration in metro optical DWDM networks is the total transmission length and number of cascaded nodes that can be satisfied by system optics without resorting to the cost and complexity of electrical regeneration. The use of cost-effective devices (e.g., directly modulated transmitters) to reduce the overall cost the need for network reconfiguration render the engineering of a transparent metro network a nontrivial task. Special care should be taken to reduce the degradation of the signal quality due to the use of such cost-effective technologies. Moreover, depending on the network and network node architecture, the impact of some of the effects can be more pronounced. In the following paragraphs we discuss the main effects that make the design of system offerings for metro networks not a trivial task.

Signal attenuation from fiber/component loss could be overcome using optical amplifiers. The *noise* that these amplifiers introduce in metropolitan area networks can be managed since the size of the network is relatively small and short amplifier spans are used [high optical signalto-noise ratio (SNR) can be preserved]. Power divergence among the WDM channels caused by component ripple as well as polarization-depended loss/gain effects can degrade system performance, but they can be combated using static/dynamic spectral equalization [10,15]. Signal transients occur in metro networks because of the increased number of channel add/drops and in the case of protection switching, but the use of dynamic gain-controlled amplifiers and loss-controlled attenuators can reduce their impact [10,20,27]. Filter concatenationinduced distortion is a more severe problem since, in most cases, it cannot be compensated [28,29]. In-band *crosstalk* might also limit the size of network that can be supported. However, proper network design can reduce crosstalk-induced limitations [10,15]. Polarization mode dispersion is not considered to be a severe degradation, due to the low bit rates used in metropolitan-area networks [1]. Fiber nonlinearities are seldom probed in metro networks since the channel spacing is large and the injected power per channel in the fiber can remain low because of the small amplifier spacing [1]. Chromatic dispersion is another degrading effect in metro systems, especially with the use of low-cost transmitters (e.g., directly modulated lasers/DMLs and electroabsorption modulator-integrated DFB lasers/EA-DFBs) [32]. These lasers present highfrequency chirp (i.e., the optical frequency of the emitted signal varies rapidly with time depending on the changes of the optical powers) [1], which limits the uncompensated reach [32]. The dispersion/chirp impairment can be overcome by using dispersion-compensating modules or properly engineered optical fibers with special dispersion characteristics [16,32].

Some of the abovementioned degrading effects could be assigned to the transmission line (fiber) and others to the network node equipment. It is then critical to optimize the design of the transmission fiber, the network node architecture and the equipment used in the nodes in order to achieve optimum performance. In the following, we will see the specific desirable features that the metro network requirements impose to equipment (e.g., amplifiers), fibers and OADMs.

4.1. Metropolitan Amplifiers

The ever-changing traffic demands lead to the requirement for changes in the total number of WDM channels and also in the percentage of channels added or dropped at the OADM network nodes. Such dynamic network reconfigurations will more likely occur more frequently in metropolitan than in long-haul networks. Furthermore, protection and restoration mechanisms may force the termination of some channels or traffic in the network. The discussion above indicates that the amplifiers designed for metropolitan networks should have some unique requirements. The amplifiers have to deal with very dynamic networks where a large number of channels can be added and dropped, which means that amplifiers have to respond rapidly to these events by keeping their gain, and consequently the per channel power at the amplifier output, at a constant level independent of the number of wavelengths present in the network. Another requirement for metro amplifiers is operation over a wide gain range, since they have to deal with extremely wide variations in span length and node losses.

Consequently, key features for metro amplifiers are (1) variable-gain operation and (2) fast transient gain control. Such amplifiers will enable fast provisioning in dynamically reconfigurable networks, optical protection switching, and either longer system reach or more OADM nodes per ring, or larger percentage of add/drop capability per OADM node. In amplified optical networks, the use of erbium-doped fiber amplifiers (EDFAs) with dynamic gain-control capability [35] can significantly reduce the signal transients, and allow for tunable gain characteristics. Other amplifiers (SOAs) have shown potential for use in metro networks [36].

4.2. MAN-Optimized Optical Fibers

Depending on the size, capacity, application, and terminal equipment used in metro networks, several different fiber types could be considered. Although advanced WDM metropolitan networks borrow heavily from concepts and practices originally developed and evaluated for their longhaul cousins, there are significant differences between the two application spaces, as we have pointed out. Among these is the necessity to support a broad spectrum of services and data rates in the MAN while keeping costs as low as possible. More specifically, metro fibers need to support:

• 1310-nm operation—since the majority of legacy SONET systems operate at 1310 nm, the fiber type

2870 WDM METROPOLITAN-AREA OPTICAL NETWORKS

used in metro networks should be able to operate at that wavelength range.

- Full-spectrum (1260-1630-nm) coarse WDM (CWDM)—CWDM continues to foster growing interest as a cost-effective means of enabling efficient bandwidth usage without the complexity and tight tolerances on optics associated with DWDM systems. The ability to have future CWDM compatibility built into metro networks reflects intelligent and proactive planning.
- Large uncompensated / unregenerated reach for DWDM systems at 1550 nm—long uncompensated/unregenerated reach with low-cost transmitters will enable the reduction of OEO regenerators and dispersion compensation modules and consequently will minimize the cost of the network.

Only selected fiber types can meet all the requirements for a metro fiber. For example, the most widely deployed single-mode fiber type (standard ITU G.652 fiber) does not satisfy all the requirements. Such fiber type has attenuation and dispersion characteristics, illustrated in Fig 6. Early versions of this fiber type may not be used for CWDM because of excess attenuation around 1380 nm. For networks requiring wavelength-band service differentiation and ability for CWDM, a fiber type with reduced attenuation at that wavelength range (e.g., Corning's SMF-28e fiber, OFS's Allwave fiber) should be the optimum choice.

A fiber with reduced water peak attenuation and optimized dispersion characteristics across the available fiber bandwidth should enable the use of cost-effective transmitters while ensuring compatibility with both CWDM and DWDM implementations. For example, the use of fibers with negative dispersion at the operating wavelength (Fig. 6b) enables the deployment of cost-effective directly modulated 10-Gbps transmitters for uncompensated reach comparable to that achieved by the best externally modulated sources over standard single-mode fiber [31]. Such a choice may enable cost-effective long-reach 10-Gigabit Ethernet transport solutions.

Nonzero dispersion-shifted fibers (NZ-DSF) also have great potential for use in metro networks. The lower absolute value of dispersion and the low attenuation across the operating wavelength window (see Fig. 6) will enable long uncompensated reach, legacy SONET 1310-nm operation, and simultaneous DWDM and CWDM operation.

4.3. OADM Architectures

Currently most of the metro-system integrators are building equipment that will support WDM OADM rings.



Figure 6. (a) Attenuation and (b) Dispersion characteristics of nonzero dispersion-shifted fiber, negative-dispersion fiber, and low-water peak fiber.

Because of the dynamic nature of data traffic originating in metro networks, OADM architectures should be designed for capability of large percentage of traffic add/drop at the network nodes. Although only a few channels may actually be added or dropped at each OADM, access to a large number of channels will allow for mesh connectivity of many nodes and will help avoid wavelength blocking issues with minimum pre-planning. Remote reconfiguration of the OADMs is also desirable [8].

Typically, OADM metro ring networks are built according to a banded wavelength channel plan [10,12]. The wavelengths are split into individual bands (typically 3-8 wavelengths each), with a guard-band spacing between the wavelength bands. The channels within the band can be spaced at frequencies of 25, 50, 100, or 200 GHz. The use of wavelength banding enables hierarchical multiplexing/demultiplexing at each network node, optical node bypassing on a per band basis, and scalable capacity upgrades and consequently reduced first installed costs [12]. The idea of node bypassing is very attractive since demultiplexing every wavelength at each node can be expensive, especially with many wavelengths per fiber and possibly many fibers per node. In addition to cost reduction of switching equipment, the wavelength banding will result in improved optical transmission performance since the insertion loss at each node for the passthrough channels will be reduced and also less filter concatenation effects are expected. The wavelength bandlayered architecture consists of wavelength-band filters in combination with single-channel filters, and proper amplification to compensate for OADM losses (Fig. 7a). After the signals enter the OADM, some wavelength bands can be dropped using band-drop filters. The other bands (passthrough) experience a small loss by the band filters as they pass transparently through the OADM. The channels corresponding to the dropped bands are demultiplexed using single-channel demultiplexers and are directed to the receivers. In the add path a combination of single-channel multiplexers and band-add filters is used to add new traffic to the available channel slots. Variable optical attenuators (not shown in Fig. 7) can be used to match the power of the added channels to that of the passthrough channels. Optical amplifiers are placed at the input and the output of the OADM to compensate for losses. Wavelength-band layered OADM architectures meet most of the requirements of metro networks. However, when using such OADM architecture once the network is provisioned, the configuration is fixed, which induces limitation on the network flexibility. Furthermore, the wavelength-band layered OADM structure limits the maximum number of network nodes that can be supported for mesh connectivity.

More recently, a "broadcast & select" OADM architecture (Fig. 7b) has been proposed as an alternative to OADM applications where a large number of channels must be accessed [8,20]. The B&S OADM architecture consists of a 1×1 wavelength-selective device [e.g., Corning's PurePath dynamic spectral equalizer (DSE)] in combination with 1×2 power splitters/combiners to perform traffic add/drop, and proper amplification to compensate



Figure 7. (a) Wavelength band layered and (b) broadcast & select OADM architectures.

for OADM losses. In this architecture (Fig. 7b), all incoming traffic is split into two paths for drop and passthrough. In the drop path, the dropped traffic is selected by a combination of a power splitter $(1 \times N, \text{where } N \text{ is the number of }$ simultaneously accessible DWDM channels) and tunable filters. In the passthrough path, the dropped channels are blocked by the DSE and the available channel slots can be filled by signals coming from the add path. The add path consists of N tunable transmitters and a $N \times 1$ power combiner. An EDFA is used to compensate for losses in the add path, and the amplifier noise in the empty channel slots is then filtered by another DSE. Both DSEs in this architecture can perform signal blocking for some of the channels and simultaneous power leveling for the pass through and added traffic. EDFAs are placed at the input and the output of the OADM to compensate for OADM losses. Such architecture has the advantages of keeping the passthrough loss to a minimum, while adding flexibility to the number of channels to be accessed. The B&S OADM architecture allows an arbitrary number of add/drop channels at each node and is dynamically reconfigurable [8].

5. SUMMARY

We have presented the characteristics, requirements, and architectures of optical metropolitan-area networks. The main features of system offerings proposed for metro networks were also discussed. We outlined our considerations regarding the evolution of metropolitanarea networks from SONET rings, to transparent wavelength-routed WDM rings. We also described the major transport impairments that limit the performance of transparent WDM metropolitan-area networks, and we presented the characteristics of MAN-optimized optical amplifiers, fibers, and OADMs. More recent research results demonstrate that a properly engineered network, utilizing application-optimized components and fiber, will enable the buildup of cost-effective metrooptimized WDM optical networks that satisfy all the requirements and demands.

BIOGRAPHY

Ioannis Tomkos received the B.Sc. degree in Physics from University of Patras, Greece, and the M.Sc. degree in Telecommunications Engineering and Ph.D. degree in Optical Telecommunications from the University of Athens, Greece. His Ph.D. work focused on novel all-optical wavelength conversion technologies.

In 1996, he joined the Optical Communications Group of the University of Athens, Greece, as a Research Fellow, where he researched technologies for all-optical networks and digital transmission systems for access networks.

In January 2000, he joined the Photonics Research and Test Center of Corning Inc. as a Senior Research Scientist. He studied extensively the performance and design issues of metropolitan-area optical networks and successfully led several related projects.

Since September 2002, he has been an Associate Professor at the Athens Information Technology Center of Excellence in Research and Graduate Education, Athens, Greece. He is also an Adjunct Associate Professor at Carnegie-Mellon University, Pennsylvania (USA).

Professor Tomkos received the Best Paper Award from IEEE-LEOS in 1998. In 2002, he received the 2001 Corning Research Outstanding Publication Award. He has co-authored about 70 contributed and invited papers, has published in international journals and conference proceedings, and has several patent applications pending. He is a member of IEEE-LEOS and a member of the Optical Fiber Communication Conference (OFC) Technical Program Committee.

BIBLIOGRAPHY

- G. P. Agrawal, Fiber Optic Communication Systems, Wiley, New York, 1992.
- C. A. Brackett et al., A scalable multiwavelength multihop optical network: A proposal for research on all-optical networks, *J. Lightwave Technol.* 11(5): 736–753 (May–June 1993).
- S. B. Alexander et al., A precompetitive consortium on wideband all-optical networks, J. Lightwave Technol. 11(5): 714-735 (May-June 1993).
- R. E. Wagner, R. C. Alferness, A. M. Saleh, and M. S. Goodman, MONET: Multiwavelength optical networking, *J. Light*wave Technol. 14(6): 1349–1355 (June 1996).
- 5. T. E. Stern and K. Bala, *Multiwavelength Optical Networks*, Prentice-Hall, Englewood Cliffs, NJ, 2000.
- B. Mukherjee, Optical Communication Networking, McGraw-Hill, New York, 1997.
- N. V. Srinivasan, Add-drop multiplexers and cross-connects for multiwavelength optical networking, *Proc. Optical Fiber Communication Conf.*, OFC '98, 1998, 57–58.
- 8. A. Boscovic, M. Sharma, N. Antoniades, and M. Lee, Broadcast and select OADM nodes: Application and performance

trade-offs, Proc. Optical Fiber Communication Conf., OFC'02, Paper TuX2, 2002, 158–159.

- I. Tomkos et al., 80 × 10.7 Gb/s ultra-long-haul (4200 km) DWDM network with dynamic operation of broadcast & select reconfigurable OADMs, Proc. Optical Fiber Communication Conf., OFC'02, Paper FC-1, 2002.
- I. Tomkos et al., Transport performance of an 80 Gb/s WDM ring network utilizing directly modulated lasers and uncompensated negative dispersion fiber, *IEEE/OSA J. Lightwave Technol.* 20(4): 562-573 (April 2002).
- M. D. Vaughn and R. E. Wagner, Metropolitan network traffic demand study, Proc. Lasers and Electro-Optics Society IEEE Annual Meeting, LEOS 2000, 2000, Vol. 1, pp. 102–103.
- A. A. M. Saleh and J. M. Simmons, Architectural principles of optical regional and metropolitan access networks, J. Lightwave Technol. 17(12): 2431–2448 (Dec. 1999).
- D. Stoll, P. Leisching, H. Bock, and A. Richter, Metropolitan DWDM: A dynamically configurable ring for the KomNet field trial in Berlin, *IEEE Commun. Mag.* **39**(2): 106–113 (Feb. 2001).
- P. Leisching et al., All-optical-networking at 0.8 Tb/s using reconfigurable optical add/drop multiplexers, *IEEE Photon. Technol. Lett.* 12(7): 918-920 (July 2000).
- N. Antoniades et al., Performance engineering and topological design of metro WDM optical networks using computer simulation, *IEEE J. Select. Areas Commun.* (Special Issue on WDM Based Network Architectures) **20**(1): 149–165 (Jan. 2002).
- J.-P. Faure et al., A scalable transparent waveband-based optical metropolitan network, *Proc. Eur. Conf. Optical Communication, ECOC '01*, 2001, Vol. 6, 64–65.
- K. C. Reichmann et al., An eight-wavelength 160-km transparent metro WDM ring network featuring cascaded erbium-doped waveguide amplifiers, *IEEE Photon. Technol. Lett.* 13(10): 1130-1132 (Oct. 2001).
- P. Arijs et al., Architecture and design of optical channel protected ring networks, J. Lightwave Technol. 19(1): 11-22 (Jan. 2001).
- D. Tebben et al., Two-fiber optical shared protection ring with bi-directional 4/spl times/4 optical switch fabrics, Proc. Annual Meeting of Lasers and Electro-Optics Society, LEOS 2001, 2001, Vol. 1, pp. 228-229.
- 20. J.-K. Rhee et al., A novel 240-Gbps channel-by-channel dedicated optical protection ring network using wavelength selective switches, *Proc. Optical Fiber Communication Conf.*, *OFC'01*, Paper PD-38, 2001.
- P. Bonenfant and A. Rodriguez-Moral, Framing techniques for IP over fiber, *IEEE Network* 15(4): 12–18 (July–Aug. 2001).
- 22. Yao Shun, S. J. B. Yoo, B. Mukherjee, and S. Dixit, Alloptical packet switching for metropolitan area networks: Opportunities and challenges, *IEEE Commun. Mag.* **39**(3): 142–148 (March 2001).
- 23. K. V. Shrikhande et al., HORNET: A packet-over-WDM multiple access metropolitan area ring network, *IEEE J. Select. Areas Commun.* 18(10): 2004–2016 (Oct. 2000).
- A. Jourdan et al., The perspective of optical packet switching in IP dominant backbone and metropolitan networks, *IEEE Commun. Mag.* **39**(3): 136–141 (March 2001).

- N. Le Sauze et al., A novel, low cost optical packet metropolitan ring architecture, Proc. Eur. Conf. Optical Communication, ECOC '01, 2001, Vol. 6, pp. 66–67.
- N. Antoniades, M. Yadlowsky, and V. L. daSilva, Computer simulation of a metro WDM interconnected ring network, *IEEE Photon. Technol. Lett.* **12**(11): 1576-1578 (Nov. 2000).
- 27. J.-K. Rhee, I. Tomkos, P. Iydroose, and M.-J. Li, Optical transient effect of dedicated optical channel protection in a two fiber ring network, *Proc. Lasers and Electro-Optics Society IEEE Annual Meeting, LEOS 2001*, 2001, Vol. 1, pp. 226–227.
- J. Downie et al., Effects of filter concatenation for directly modulated transmission lasers at 2.5 and 10 Gb/s, *IEEE/ LEOS J. Lightwave Technol.* 20(2): 218-228 (Feb. 2002).
- I. Tomkos et al., Filter concatenation penalties for 10 Gb/s sources suitable for WDM metropolitan area networks, *IEEE Photon. Technol. Lett.* 14(4): 564–566 (April 2002).
- C.-C. Wang et al., Negative dispersion fibers for uncompensated metropolitan networks, *Proc. Eur. Conf. Optical Communication, ECOC'00*, Munich, Germany, Sept. 2000, Vol. 1, pp. 70–71.
- 31. I. Tomkos, R. Hesse, R. Vodhanel, and A. Boskovic, A 320 Gb/s metropolitan area ring network utilizing 10 Gb/s directly modulated lasers, *IEEE Photon. Technol. Lett.* 14(3): 408-410 (March 2002).
- 32. I. Tomkos et al., Demonstration of negative dispersion fibers for DWDM metropolitan area networks, *IEEE/LEOS J. Select. Top. Quant. Electron.* (Special Issue on Specialty Fibers) 7(3): 439-460 (May-June 2001).
- 33. A. Richter et al., Optical performance monitoring in transparent and configurable DWDM networks, *IEE Proc. Optoelectron.* 149(1): 1–5 (Feb. 2002).
- E. Park, Error monitoring for optical metropolitan network services, *IEEE Commun. Mag.* 40(2): 104–109 (Feb. 2002).
- M. Lelic et al., Smart EDFA with embedded control, Proc. Annual Meeting of the Lasers and Electro-Optics Society, LEOS 2001, 2001, Vol. 2, pp. 419-420.
- D. A. Francis, S. P. DiJaili, and J. D. Walker, A single-chip linear optical amplifier, *Proc. Optical Fiber Communication Conf., OFC 2001*, 2001, Vol. 4, pp. PD13, pp. 1–3.

WIDEBAND CDMA IN THIRD-GENERATION CELLULAR COMMUNICATION SYSTEMS

JON W. MARK University of Waterloo Waterloo, Ontario, Canada

SHIHUA ZHU Xian Jiaotong University Xian, Shaanxi People's Republic of China

1. INTRODUCTION

The second-generation (2G) CDMA standard, IS-95, is a narrowband CDMA standard. The third-generation (3G) CDMA systems will have a target transmission rate of 2 Mbps (megabits per second). ITU (International Telecommunications Union), the international standards body, has adopted both UMTS/IMT-2000 (Universal Mobile Telecommunications System/International Mobile Telecommunications by the year 2000) and CDMA-2000 as 3G network access technologies.

The impetus behind the popularization of wireless communications is the flexibility it offers for mobile users to roam. For wireless systems to be economically feasible, they must be able to deploy low-power transmitters and offer high system capacity, that is, be able to support a large population with a high transmission rate. These are the reasons why radio cells in wireless systems are relatively small (e.g., microcells) and arranged as a cellular structure. Communications by mobile users in a cellular environment encounters a number of challenging problems. These include the need to (1) expand the spectral width of the transmission channel, (2) manage the available resources efficiently, and (3) manage the user mobility effectively and efficiently. The 3G standards have specifications to address the above-mentioned problems. Also, in CDMA systems, there is a need for power control, at least to combat the near-far effects [6,7,11]. Because of space limitation, in this article we mainly focus attention on the network architecture and signaling strategies of IMT-2000 wideband CDMA (WCDMA).

2. NETWORK ARCHITECTURE OF IMT-2000

The UMTS/IMT-2000 architecture is shown in Fig. 1. The symbols used in Fig. 1 are listed in Table 1. The core network (CN) includes at least the circuit-switched mobile switching center (MSC) and the packet-switched packet radio service support node [2].

2.1. Functional Characteristics

In the UMTS/IMT-2000 system, the functional layering introduces the concepts of access stratum and nonaccess stratum. All the functional blocks in the UTRAN belong to the access stratum. The UTRAN handles all radio-specific procedures, whereas the core network handles the servicespecific procedures, including mobility management and call control. The other functional characteristics are as follows:

- The link Uu (see Fig. 1) connects the two physical layers (UE and UTRAN) to provide a transparent digital signal path to the upper layers. The physical layer occupies certain bandwidth, and is responsible for synchronization.
- The MAC layer resolves the contention resulting from the radioband sharing between a number of users. It also maps the data from the link control layer to the physical layer, and vice versa.
- The RLC provides reliable data packet transmission over the radio link. The signaling from the control plane can be regarded as a special data packet requiring higher priority and short delay.
- The LAC provides protection to the user data.
- The RRC performs the radio resource management, and allocates the transmission speed and power for each connection.



Figure 1. UMTS/IMT2000 system architecture and protocol layering.

Table 1. Definitions of Symbols Used in Fig. 1

Symbol	Description			
UE	User equipment			
UTRAN	UMTS terrestrial radio-access network			
CN	Core network			
Uu	Radio interface between UTRAN and UE			
Iu	Interface between UTRAN and CN			
CPS	C plane, signaling			
UPD	U plane, data			
UPS	U plane, speech			
$\mathbf{C}\mathbf{M}$	Connection management			
$\mathbf{M}\mathbf{M}$	Mobility management			
LAC	Link-access control			
RLC-C	Radio-link control for the control plane			
RLC-U	Radio-link control for the user plane			
RRC	Radio resource management and rate allocation			
MAC	Medium-access control			

- The MM maintains a record of the positions of all the users in the system and provides the updated user position information during a call.
- The CM grants or rejects an incoming call, either newly generated or handed over from an adjacent cell, based on the channel resources available at the time of call arrival.

3. PRINCIPLE OF CDMA

3.1. Orthogonal Spreading

It is well known [9] that in CDMA, if the spreading signals, $\phi_i(t) \in L^2(T), i = 1, 2, ..., M$, are orthogonal:

$$\int_{0}^{T} \phi_{k}(t)\phi_{l}(t) dt = \delta_{kl} \stackrel{\Delta}{=} \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$$
(1)

where *T* is the signal duration; the spread signals $x_i\phi_i(t)$, i = 1, 2, ..., M, will also be orthogonal to each other. Here x_i is the data bit of the *i*th user, which is kept constant during *T*. Consequently, the interference from any other

signal, say, $x_j\phi_j(t)$, on the wanted signal, say, $x_i\phi_i(t)$, is theoretically zero after despreading:

$$I_{ij} = \int_0^T x_j \phi_j(t) \phi_i(t) \, dt = x_j \int_0^T \phi_j(t) \phi_i(t) \, dt = 0, \quad i \neq j \quad (2)$$

where $\phi_i(t)$ is the despreading function used to extract the desired data sequence, $\{x_i\}$, transmitted by the *i*th user.

It is not difficult to design a set of orthogonal functions or codes. Some good examples are Hadamard–Walsh codes [4]. However, when these codes are not properly aligned, the cross-correlation of these codes will be nonzero, and may be relatively large compared to the autocorrelation; that is, the signals are no longer orthogonal.

Thus, to eliminate the interference from unwanted signals, the local despreading code is required to be accurately aligned with the arriving wanted code. This demands that all the signals arriving at every receiver be synchronized at the bit level and with the spreading code. In other words, if it is desired to avoid interferences from other users using the orthogonality property, all the received signals must be bit-synchronized with the local timing that is locked to the spreading code of the incoming wanted signal. For practical systems in which mobile terminals are constantly in motion, this is obviously a very costly requirement. It is desirable to have a type of code that can separate channels and yet requires no synchronization.

To date, no orthogonal code requiring no synchronization has been found. In frequency diversity methods such as FDMA, if a certain amount of interference from adjacent bands is tolerable, a less sharp cutoff filter can be employed, which can greatly simplify the filter design. Similarly, in CDMA if nonperfect channelization is tolerable, nonorthogonal codes that require no synchronization may be used.

3.2. Nonorthogonal Spreading

The nonorthogonal codes when used for spreading must require no synchronization at the bit level and with the spreading code. To reduce the interference from other transmissions to a minimum, the cross-correlation, Γ_c , between any pair of codes of the code set at any time shift should be small; thus, we require

$$\Gamma_c = \int_0^T \phi_k(t) [\phi_l(t - T + \tau) + \phi_l(t + \tau)] dt \ll 1 \quad 0 \le \tau < T$$
(3)

Besides, for reasons to be explained in the next section, the codes are required to be balanced; that is, the number of ones must approximately equal the number of zeros. Gold codes and Kasami codes have such properties [4,5,10]. Kasami codes consist of two classes: the large Kasami code sets and the small Kasami code sets. The large Kasami code sets are of more interest to 3G CDMA applications.

3.3. Two-Layer Spreading

In two-layer spreading, the transmitter signal is spread in the first layer using orthogonal codes. The output from the first layer is then spread again using nonorthogonal codes. Orthogonal codes are used in the first layer spreading for reasons that synchronization between channels can be easily maintained. Nonorthogonal codes are used for the second layer to reduce implementation complexity.

Two-layer spreading is illustrated in Fig. 2, where the set of parallel signals, $\{x_i(t), i = 1, 2, ..., K\}$, belongs to one transmitter. All the channels originating from a single transmitter can be readily synchronized. Therefore, orthogonal codes $(C_{o1}, C_{o2}, ..., C_{oK})$, are used to separate these channels. These channels are then linearly combined and multiplied by a transmitter-specific nonorthogonal code.



Figure 2. Two-layer spreading.

In two-layer spreading, every transmitter uses the same orthogonal code set. Let the data signal on the kth channel from the mth transmitter be represented by

$$x_{mk}(t) = \sum_{i} b_{mki} g_{T_b}(t - iT_b) \tag{4}$$

where T_b is the bit duration, $b_{mki} \in \{-1, 1\}$ is the *i*th bit of the *k*th channel from the *m*th transmitter, and $g_{T_b}(t)$ is a rectangular pulse of duration T_b :

$$g_{T_b}(t) = \begin{cases} 1 & 0 \le t < T_b \\ 0 & \text{elsewhere} \end{cases}$$
(5)

Let $c_{okj} \in \{-1, 1\}$ be the *j*th chip of the *k*th orthogonal code. Then, the orthogonal spreading code used for the *k*th channel can be expressed as

$$c_{ok}(t) = \sum_{j=1}^{N} c_{okj} g_{T_c}(t - jT_c)$$
(6)

where T_c is the chip width and N is the code length. Normally, the values of T_b and T_c are chosen to yield $T_b/T_c = N$, resulting in the same code repeatedly multiplied onto each data bit of the channel.

The length of the nonorthogonal codes is normally many times the length of orthogonal codes. They can be a set of codes, or a set of segments of some long code. Their chip rate is the same as that of the orthogonal code: $1/T_c$. The nonorthogonal signal can be written as

$$c_{sm}(t) = \sum_{l=1}^{N'} c_{sml} g_{T_c}(t - lT_c)$$
(7)

where $c_{sml} \in \{-1, 1\}$ is the *l*th chip of the nonorthogonal code for the *m*th transmitter and N' is the code length. The output from the two-layer spreading is then given by

$$y_m(t) = \sum_{k,i} b_{mki} c_{ok}(t - iT_b) \sum_p c_{sm}(t - pT')$$
(8)

where $T' = N'T_c$ is the nonorthogonal code duration. Figure 3 shows the timing of the waveforms in a twolayer spreading system, where $x_{mk}(t)$ is the *k*th channel signal from the *m*th transmitter.



Figure 3. Timing of two-layer spreading.

In a CDMA system, many users transmit signals in the same band. The received signal (in the absence of channel impairments and background noise) is given by

$$r(t) = \sum_{m} \sum_{k,i} b_{mki} c_{ok} (t - \tau_m - iT_b) \sum_{p} c_{sm} (t - \tau_m - pT')$$
(9)

where τ_m is the transmission delay from the *m*th transmitter to the target receiver.

To receive the desired signal, say, $x_1(t)$ of transmitter 1, the receiver must be synchronized to the nonorthogonal code frame; that is, when the receiver achieves synchronization, $\tau_1 = 0$. The output from the nonorthogonal despreading is

$$z'(t) = r(t) \sum_{p} c_{s1}(t - pT')$$

= $\sum_{k,i} b_{1ki}c_{ok}(t - iT_b)$
+ $\sum_{m \neq 1} \sum_{k,i} b_{mki}c_{ok}(t - \tau_m - iT_b)$
 $\times \sum_{p} c_{sm}(t - \tau_m - pT')c_{s1}(t - pT')$ (10)

After orthogonal despreading, the output becomes

$$z(t) = z'(t) \sum_{i} c_{o1}(t - iT_{b})$$

$$= \sum_{i} b_{11i} + \sum_{k \neq 1, i} b_{1ki}c_{ok}(t - iT_{b})c_{o1}(t - iT_{b})$$

$$+ \sum_{m \neq 1} \sum_{k, i} b_{mki}c_{ok}(t - \tau_{m} - iT_{b})c_{o1}(t - iT_{b})$$

$$\cdot \sum_{p} c_{sm}(t - \tau_{m} - pT')c_{s1}(t - pT')$$
(11)

The first term on the RHS of the second equality in (11) is the wanted signal. The second term represents the interferences arising from the other channels of the same transmitter, which vanishes after integration due to the orthogonality between $c_{o1}(t - iT_b)$ and $c_{ok}(t - iT_b), k \neq 1$. The third term is the interference contribution from the other transmitters. Since the transmitters are not synchronized, the first part of the product may not be zero after integration; it may even be 1 (suppose k = 1and $\tau_m = iT$). This is equivalent to that, at any time t, the probability of the signal being 1 and being -1 is not equal. But if $c_{sm}(t - \tau_m - pT')$ and $c_{s1}(t - pT')$ have small correlation values and are balanced, namely, at any time t, the probability of the signal being 1 is approximately the same as it being -1. After multiplication, the product is approximately balanced and approaches zero after integration. This is the scrambling technique normally used in data transmission. For this reason, the nonorthogonal codes are called scrambling codes. Similarly, the orthogonal codes are called *channelization* codes.

It can be seen from Fig. 3 that when the spreading factor is small, there is a possibility that the segment of

the scrambling code corresponding to a bit interval (or a channelization code period) is far from balanced. If we assume that the number of interfering transmitters is large and the interferences are independent of each other, the statistics of the interferences can still be regarded as random noise. It is certain, however, that a smaller spreading factor will give a worse scrambling effect than will a larger spreading factor.

3.4. Spreading in IMT-2000

UMTS/IMT-2000 employs two-layer spreading. For the downlink, every base station uses the same set of Hadamard codes or OVSF (orthogonal variable spreading factor) codes as its channelization codes. Each cell uses a cell-specific Gold code as its scrambling code.

For the uplink, every active mobile UE establishes a connection with the base station. The connection may consist of one or several channels. Every UE or connection uses the same Hadamard/OVSF codes as its channelization code. Each connection is then distinguished by using a user-specific Gold code or Kasami code as its scrambling code.

Obviously, since there are many more downlink channels from a single base station than uplink channels from a single UE, a much larger set of channelization codes for the downlink transmission is needed. Similarly, since the number of UEs in a cellular system is much larger than that of base stations, a larger scrambling code set for the uplink transmission is required.

In UMTS/IMT-2000, the same OVSF codes are used for both links. Since the maximum capacity of OVSF codes has been proved to be bounded by the transmission rate, special attention must be paid to the downlink channelization code usage, so that more connections can be accommodated by the given channelization code set. This has led to different structures for the uplink and downlink physical channels (see Section 4.2).

The scrambling codes used for the downlink is a set of computer-selected 10-ms segments of a $(2^{18} - 1)$ -chip-long Gold sequence. Each segment is equivalent to 40,960 chips for the recommended 4.096-Mchip/s¹ transmission rate at a carrier spacing of 5 MHz. For the uplink, a larger number of scrambling codes is required. They can be either a set of the extended 256-chip-long large Kasami codes, or optionally, a set of 10 ms (40,960 chips) segments selected from a $(2^{41} - 1)$ -chip-long Gold sequence [4].

4. PHYSICAL CHANNELS

4.1. Physical Channel Types

UMTS/IMT-2000 defines the following physical channel types (see Fig. 1 for references to protocol layers):

¹ The originally recommended chip rate for IMT-2000 was 4.096 Mchips/s at a bandwidth of 5 MHz. In the move toward harmonized global 3G (G3G), ITU adopted a compromised chip rate of 3.84 Mchips/s for 3G. For the purpose of discussing IMT-2000 spreading in this article, we continue to use 4.096 Mchips/s as the reference. For the 3.84-Mchip/s rate, proper scaling can easily be accommodated.

- Synchronization channel (SCH) used for initial cell search and link synchronization by the UE so that the data and control channels can be properly despreaded (downlink only).
- Dedicated physical data/control channel (DPDCH/ DPCCH)—each connection between a dedicated UE and the network is allocated one DPCCH and zero, one, or several DPDCHs. The DPDCH is used to carry the data generated at layer 2 and above, dedicated to a single UE; the DPCCH is used to carry layer 1 control information relevant to the DPDCH.
- Common control physical channel (CCPCH)—used to broadcast common control information from layers 2 and 3 to all UEs of a cell or the whole system (downlink only).
- Physical random-access channel (PRACH)—provides the UEs fast access of short data packets to the network. This is a supplement to the DPDCH/DPCCH and is particularly valuable for control information and signaling transmission (uplink only).

4.2. Dedicated Physical Data/Control Channels

The *frame structures* of IMT-2000 are shown in Figs. 4 and 5, with the following parameter values:

- Data sequence is divided into frames of length $T_f = 10$ ms.
- Each frame is further divided into 16 slots of equal duration $T_s = 0.625$ ms.
- The uplink DPDCH and DPCCH each occupies a separate channel, or one frame per 10 ms. Each slot of the DPCCH is divided into three fields, for the known pilot bits, transmit-power-control (TPC) command, and transport format indicator (TFI), respectively, as shown in Fig. 4. The pilot bits are used for downlink channel estimation. The reason for a dedicated pilot instead of a common pilot is to support the use of adaptive antenna arrays. The TPC is used to perform closed-loop power control at a frequency of 1600 Hz. The TFI (optional) is used to inform the receiver which transmit format the layer 2 data is used in the current DPDCH.
- The downlink DPDCH and DPCCH are multiplexed within each radio frame as shown in Fig. 5. The main purpose of this multiplexing is to let the two channels share a single channelization code.



Figure 4. Uplink DPDCH/DPCCH frame structure.



Figure 5. Downlink DPDCH/DPCCH frame structure.

4.2.1. Variable Orthogonal Spreading

- On the uplink (Fig. 6), the allowable data bits carried by each slot are $n_{su} = 10,20,40,80,160,320,640$ $(10 \times 2^k, k = 0, 1, \ldots, 6)$ bits. These correspond to channel data rates $R_{bu} = n_{su}/T_s = 16,32,64,128$, 256,512,1024 $(16 \times 2^k, k = 0, 1, \ldots, 6)$ kbps (kilobits per second). The spreading output is kept at a constant rate $R_c = 4.096$ Mchips/s, resulting in variable spreading factors (SF) of $R_c/R_{bu} = 256,128,64,32,16,8,4$ (256/2^k, $k = 0, 1, \ldots, 6$).
- On the downlink, since the DPDCH and DPCCH are multiplexed in each slot (see Fig. 5), the number of bits carried by each slot must be doubled in order to offer the same transmission capacity for both links. Consequently, the allowable data bits carried by each slot should be $n_{sd} = 20,40,80,160,320,640,1280$ (20 × $2^k, k = 0, 1, \ldots, 6$) bits. These correspond to channel data rates $R_{bd} = n_{sd}/T_s = 32,64,128,256,512,1024,$ 2048 $(32 \times 2^k, k = 0, 1, ..., 6)$ kbps. The spreading output rate is constant: $R_c = 4.096$ Mchips/s. To obtain the same spreading factors as in the uplink, the serial datastream is first converted to two parallel sequences (see Fig. 7) before being sent for spreading. The resulting variable spreading factors are again $R_c/(R_{bd}/2) = 256,128,64,32,16,8,4$ $(256/2^k, k = 0, 1, \dots, 6).$
- The channelization codes available for each connection on the uplink and for each base station on the downlink are the same set of OVSF codes. Within an uplink connection, each DPDCH or DPCCH requires a unique channelization code for orthogonal spreading. On the downlink, each multiplexed DPDCH/DPCCH requires a unique channelization code. These codes are assigned by the network and



Figure 6. Uplink spreading and modulation.



Figure 7. Downlink spreading and modulation.

may be changed dynamically, frame by frame, during the connection.

4.2.2. Scrambling Code Spreading. For the uplink, complex (m phase and quadrature) scrambling code spreading, as shown in Fig. 6, is used. The reason for using the complex spreading is to balance the loads on the I and Q branches. To control the amount of overhead introduced by using the DPCCH, the power of the DPCCH is to be kept to a minimum. Consequently, the power of the DPCCH Besides, this difference can vary with the traffic type carried on the DPDCH. For instance, the relative power difference between the DPCCH and DPDCH for speech and 384-kbps data are 3 and 10 dB, respectively. Another advantage of this complex spreading is that additional DPDCH can be easily added to the connection by adding it to both I and Q branches after the independent channelization spreading.

The scrambling code for both I and Q branches is the same code. It can be either 256-chip-long extended codes from the VL Kasami set² of length 255, or a 40,960-chip (10-ms) segment of a Gold code of length $2^{41} - 1$ (when an ordinary RAKE receiver is used). The scrambling code on the uplink is user-specific. It is assigned by the network and may be changed dynamically during the connection.

For the downlink, because of the scarcity of channelization codes, the DPDCH and DPCCH are first multiplexed, as shown in Fig. 5, and then spread by a single channelization code, as shown in Fig. 7. The serial DPDCH/DPCCH datastream is converted into two parallel paths, I and Qpaths, so that the data rate on each path is halved, resulting in the same data rates as in the uplink case. The two branches are then spread by the same channelization code and scrambling code.

The scrambling code for the downlink is a 40,960-chip (10-ms) segment of a Gold code of length $2^{18} - 1$. There are 512 different segments, which are divided into 32 groups, each consisting of 16 codes. Each cell is assigned a specific downlink code at the initial deployment [3].

4.2.3. Filtering and Modulation. The filter p(t) is a root-raised cosine function with a rolloff factor of

 2 The VL Kasami code is simpler and the cross-correlation properties are maintained between symbols, but has worse interference averaging properties. Its use is more appropriate when multiuser detection is used. 0.22. The modulation scheme is quaternary phase shift keying (QPSK).

4.3. Synchronization Channels

To access the network, a mobile UE must be synchronized to the desired cell to detect the information. From the description in the last section, the UE must first establish the scrambling code timing in order to decode the information sent from the cell site. Since the scrambling code boundary is aligned with the bit or channelization code boundary, orthogonal despreading timing can be readily obtained, once the scrambling code timing is derived.

The most straightforward method of acquisition of the scrambling code timing is to correlate the locally generated scrambling sequence with the incoming signal. The search can be performed by shifting the local sequence phase one-half a chip at a time, until the complete period is searched. In the worst case, the mobile UE may need to search all the possible scrambling codes used by the base stations. Since the scrambling codes must be sufficient to provide cell-site separation, the number of scrambling codes is usually very large. This makes the above searching algorithm extremely time-consuming, usually too slow to be acceptable for normal operation. To cope with the difficulty, a special mechanism has been developed for UMTS/IMT-2000.

4.3.1. Synchronization Channel Types. The system employs two types of synchronization channels (SCHs): the primary SCH and the secondary SCH. The primary SCH transmits a 256-chip-long Gold sequence, called the primary synchronization code (PSC), in each slot (see Fig. 8). The sequence is system-specific and is predetermined for the whole system; that is, it is transmitted by every base station in the system. It is not spread further by any other spreading code but is superimposed on the scrambled downlink datastream. At the start of a slot, the channel transmits 0.0625 ms of this 256-chip sequence (256 bits/4096 kbps), pauses, and transmits the code again when the next slot starts. The code is transmitted timealigned with the slot boundary so that by acquiring the PSC, the mobile UE acquires slot synchronization to the target base station.

The system allocates 16 Gold codes of 256 chip length, called the *secondary synchronization code* (SSC), for indicating the 16 scrambling code groups used for downlink data and control channels (see Section 4.2.1).



Figure 8. Synchronization channel frames (*key*: C_p —primary synchronization code; C_{si} —the *i*th code of the 16 possible secondary synchronization codes; d_j —the *j*th bit of the 16-bit sequence).

Each base station is assigned a SSC by the network. The code is transmitted once per slot on its secondary SCH. As in the primary SCH, each code occupies 0.0625 ms (see Fig. 8). The SSC is further modulated by a 16-bit system-specific sequence in such a way that each 256-chip code in a slot is modulated by one bit of the 16-bit sequence, resulting in a repeat pattern every 16 slots or one frame. By correlating the local code to the SSC, the UE can determine which code group the target cell uses, and by demodulating the 16-bit sequence, the frame timing is obtained.

4.3.2. Initial Cell Search and Synchronization. The process of the initial cell search and synchronization by a mobile UE is done in three steps:

- 1. The UE receiver has a matched filter, matched to the primary SCH code, C_p . The output of the filter responding to the C_p sent from any cell is therefore a sequence of narrow peaks spaced one slot (0.625 ms) apart. Since all the cells in the system are transmitting the same C_p , the matched-filter output is a superposition of many such sequences. The UE selects the one with the largest peak, and takes the corresponding cell as its working base-station cell. This peak sequence also provides the UE with the slot timing for that cell.
- 2. With the local slot boundary aligned to the cell slot boundary obtained in step 1, the UE tries to correlate each of the 16 possible secondary SCH codes, C_{si} , i = 1, 2, ..., 16, with the received signal. The correlation is carried over a frame, and for each possible SSC the searching process is done by shifting one slot at a time. This results in a maximum of 16 correlations per code, or 256 correlations in total, for the UE to search through all the possible SSCs and the possible 16-bit sequence phases. When the code selected by the UE is matched to the secondary SCH code used by its working cell, and the local 16-bit sequence is matched to the incoming one, the search is complete. The code group that the cell uses for its downlink data and control channels is determined from the matched Gold code, while the frame timing is obtained by the 16-bit sequence phase.
- 3. Once the code group used by the cell for its downlink data and control channels is determined, the UE tries all the 32 codes within the group. (Note: These are 40,960-chip segments of a Gold code, which should not be confused with the 16 codes used by the secondary SCH, see the last section.) This is done by correlating each code with the primary CCPCH. The primary CCPCH uses a system-specific channelization code and has a fixed predefined data rate, which makes the scrambling code detection the simplest task among all the data and control channels. The search is performed through symbol-by-symbol correlation over a frame. Once the correlation operation is finished, the downlink scrambling code used by the target cell is determined and information can then be retrieved.

4.4. Common Control Physical Channels

There are two types of common control physical channels: the primary common control physical channel (primary CCPCH) and the secondary common control physical channel (secondary CCPCH). The primary CCPCH is used to broadcast system-specific information, while the secondary CCPCH is used to broadcast cell-specific information.

The framing, channelization, and scrambling for the CCPCHs are done in the same way as for the downlink time-multiplexed DPDCH/DPCCH, but with the following exceptions:

- There is no TPC field (see Fig. 5). Since it is a common point-to-multipoint broadcast channel, no closed-loop power control can be applied. Also, since both the primary and secondary CCPCHs have a fixed data rate and transport format, there is also no need for TFI. As a result, the layer 1 control information of the downlink CCPCHs (equivalent to the DPCCH of the dedicated physical channel) conveys pilot bits only.
- The primary CCPCH uses a predefined systemspecific channelization code of length 256 chips common to all cells. The secondary CCPCH uses a cell-specific code. A cell can have one or several secondary CCPCHs. The data rate for each channel is fixed during the transmission but may vary for different secondary CCPCHs within the cell and between cells. The channelization code used and the data rate carried by each secondary CCPCH is indicated by the information broadcast on the primary CCPCH (instead of on the TFI field of the layer 1 control information channel).

4.5. Physical Random-Access Channels

The physical random-access channel (PRACH) is used to carry random-access bursts and short packets in the uplink [2]. It is a common channel shared among all users in the cell. The random-access burst structure of the PRACH is shown in Fig. 9.

The preamble part is a 16-bit word spread by a cellspecific 256-chip-long Gold code that is indicated on the primary CCPCH. The rest is called the data part of the burst consisting of a UE ID, a "requested service" field, an optional user data packet, and a CRC field. Spreading and modulation of the data part is similar to the uplink dedicated channels. (So the length of the whole burst is preample $+n \times 10$ ms = $16 \times 256/4096 + 10n$ ms = 1 +10n ms.) The operation of the PRACH will be described in Section 5.3.

5. TRANSPORT CHANNELS

5.1. Transport Channel Types

The physical layer offers information transfer services to the MAC layer. These services are accomplished by

Preamble	Mobile station ID	Requested service	User packet	CRC
----------	----------------------	-------------------	-------------	-----

Figure 9. Ransom-access burst structure of the PRACH.

conveying MAC layer channels on the physical channels. These MAC layer channels, including those listed below, are denoted as transport channels (TrCh's).

- *Broadcast channel* (BCCH)—used for downlink only. With a fixed rate, it is used to broadcast system information to all the mobile UEs in its cell.
- *Paging channel* (PCH)—used for downlink only; also used for paging in the whole cell.
- *Forward-access channel* (FACH)—employed for downlink only. It is used to convey data to one or more UEs, which may be over a part of the cell using beamforming.
- *Random-access channel* (RACH)—used for uplink only. It is used by the UE to transmit short user data packets and control packets (e.g., for initiating packet transfer on the DCHs).



Figure 10. Mapping of transport channels to physical channels in IMT-2000.

• *Dedicated channel* (DCH)—a bidirectional channel. It is a point-to-point channel used to convey data to/from a UE. Beamforming can be used to achieve this.

5.2. Mapping of Transport Channels to Physical Channels

5.2.1. Mapping Relation. The mapping relation of the transport channels and the physical channels is shown in Fig. 10. The SCH is provided solely for physical layer use. One or more DCHs can be mapped onto a DPDCH/DPCCH pair. Two multiplexing schemes are proposed in Sections 5.2.2 and 5.2.3.

5.2.2. Coded Composite TrCh (CC-TrCh). In this scheme (see Fig. 11) several TrCh's are coded and interleaved individually and then multiplexed to form a coded composite TrCh (CC-TrCh). The physical layer allocates a DPDCH for the CC-TrCh and generates additionally an associated DPCCH for the connection.

5.2.3. Code-Multiplexed TrCh. In this scheme (see Fig. 12) each TrCh is coded and interleaved, and the result is sent to the physical layer which occupies a single DPDCH. So multiple DPDCHs and a single DPCCH will be generated at the physical layer for the connection.

Obviously, the coded composite TrCh is more efficient in channelization code usage but has poorer performance because it produces a high data rate, which results in a smaller spreading factor. As explained before, it is more suitable for downlink transmission. In contrast, the code multiplexed TrCh is less efficient in channelization code usage but has better performance, so it is more favorable in uplink transmission.

The processing of each block in Figs. 11 and 12 is as follows [2]:

1. *Channel Coding and Interleaving*. Depending on the specific requirements in terms of error rates, delay,



Figure 11. Coded composite TrCh in the IMT-2000.



Figure 12. Code-multiplexed TrCh's in the IMT-2000.

and other factors, the coding can take one of the following forms:

Rate- $\frac{1}{3}$ convolutional coding for low-delay services with moderate error-rate requirements, such as voice.

A concatenation of rate- $\frac{1}{3}$ convolutional coding and outer Reed–Solomon coding + interleaving for highquality services, such as data.

Turbo codes for high-rate high-quality services, such as data.

2. *Rate Matching*. Rate matching, which can be fixed or dynamic, is used to match the bit rate of the CC-TrCh to one of the available bit rates of the physical channels.

Static rate matching is carried out only when we add, remove, or redefine a TrCh. It is applied after channel coding, and uses code puncturing (decreasing rate) to adjust either the uplink or downlink rate to the physical channel rate in such a way that approximately the same SIR is achieved. But on the downlink, the rate should be adjusted to the closest lower physical channel rate to avoid the overallocation of the orthogonal codes.

Dynamic rate matching is carried out every 10 ms to match the TrCh rate to the physical channel rate by symbol repetition (increasing rate). It applies to the *uplink* only. For the downlink, discontinuous transmission within each slot is used when the instantaneous rate of the CC-TrCh does not exactly match the physical channel rate. When the data to be sent in a slot are sent off, the transmission stops until the next slot starts.

The parameters of all these processings are contained in the TFI field of each slot of the DPCCH. All the preceding processings are performed at the physical layer, but under the control of the radio-resource controller at the RRC layer (see Fig. 1).

5.3. Random Access

To facilitate burst data packet transmission, The UMTS/IMT-2000 has special arrangement for randomaccess capability.

5.3.1. Random-Access Procedure. The random-access procedure is based on slotted ALOHA and works as follows:

- 1. The UE acquires chip and frame synchronization with the target cell using the initial cell-search procedure described in Section 4.3.
- 2. The BCCH is read to retrieve information about the random-access scramble code and channelization code(s) used in the target cell.
- 3. The downlink path loss is estimated from the pilot bits of the primary CCPCH (see Section 4.4). The result is used to calculate the required transmit power of the random-access burst.
- 4. A random-access burst is transmitted on the RACH with a random time offset. The time offset is a

multiple of 1.25 ms relative to the received frame boundary.

- 5. The base station responds with an acknowledgment on the FACH.
- 6. If the UE receives no acknowledgement, it selects a new time offset and tries again.

5.3.2. Random-Access Services. There are three different service classes:

- 1. *Packet Data Services*. The packet data can be transmitted in three different ways:
 - a. If a small amount of data is to be sent, the data is simply appended to the access burst (see Fig. 13).
 - b. If the data packet is large, the access burst is sent on the RACH while the data are transmitted on the DCH. The UE first sends a "resource request" (Res_Req) message on the RACH, which indicates the message type to be transmitted. If the network has the necessary resource, it transmits on the FACH a "resource allocation" (Res_All) message which contains a set of TFs. Exactly which TF the UE may use on the DCH and at what time the UE may initiate its transmission is indicated by a "capacity allocation" (Cap_All) message, which may be sent together with the Res_All message (if traffic is light), or in a separate Cap_All message at a later time (if traffic is heavy). The TF can be changed within the given TF set during the connection and this change is indicated by a TF_Change message, which contains the new TF to be used, on the DCH.
 - c. A third method is used when the UE already has a dedicated channel at its disposal. In case when the UE has only a small amount of data to transmit, it can just start transmitting. If the UE has a large amount of data to transmit, it sends a Cap_Req message on the DCH before transmitting the data.
- 2. *Real-Time Services*. The real-time data transmission is very similar to the second way of packet data transmission, but with the following exceptions:
 - a. The UE starts transmitting immediately after the Res_All message is received. In contrast, in the case of packet data, a further Cap_All message must be received before the UE can start transmission.
 - b. Any TF in the TF set allocated in the Res_All message is allowed to be used by the UE. This makes the UE capable of supporting variable bit rate services.
 - c. The TF set can be limited to a smaller subset by a "resource limit" (Res_Limit) message if the

Arbitrary							
Access burst	Data	time	Access burst	Data			

Figure 13. Packet random access on the RACH.

network resources are insufficient. The set can be made fully available again when later the network resources become sufficient.

3. *Mixed Services*. For the mixed services, such as data and voice, two TF sets are assigned to the UE. The UE uses the voice TF set the same way as in the single-service case, while the data TF is adapted to the voice TF usage in such a way that the aggregate output power and rate should never exceed the threshold set by the MAC.

6. SUMMARY

Wideband CDMA is a de facto air interface for thirdgeneration mobile communications systems. The IMT-2000 WCDMA adopted by ITU has many salient features:

- 1. It employs a two-layer spreading; no synchronization is required between different transmitters.
- 2. It provides a wide range of services such as voice, video, data, image, and multimedia. The bit rate per channel can vary from a few kbps to 2 Mbps, either constant bit rate or variable bit rate. The information can be real-time or non-real-time. The transfer mode can be circuit-, packet-, STM-, or ATM-switched.
- 3. It provides high quality of services: toll quality for speech; BER less than 10^{-6} for data; low packet loss, delay, and delay jitter; soft handoff and macro diversity performance. Physical randomaccess channels are provided for fast access of short data packets from UE to the network.
- 4. It is flexible in system deployment and service provision, requiring simplified frequency planning; allowing macro-, micro-, picocell and employing advanced approaches of soft handoff.

In this article, because of space limitation we have focused attention mainly on signal transmission and reception strategies in the physical layer. Managing radio resources, handoff procedures, call admission control, and so on in the link layer is equally important for 3G deployment. IMT-2000 also has specifications in place to address these issues.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Grant no. RGPIN7779.

BIOGRAPHIES

Jon W. Mark received his Ph.D. degree in electrical engineering from McMaster University, Hamilton, Ontario, Canada, in 1970.

He was a professor of electrical and computer engineering at the University of Waterloo from 1970 to 2001. He is currently a distinguished professor emeritus and director of the Centre for Wireless Communications at the University of Waterloo. He had previously been at Westinghouse Canada Ltd.; IBM Thomas J. Watson Research Center, USA; Bell Laboratories, USA; Universite Pierre et Marie Curie, Paris, France; and the National University of Singapore.

He had previously worked in the areas of sonar signal processing, adaptive equalization, image and video coding, spread spectrum communications, computer communication networks, ATM switch design, and traffic management. His current research interests are in broadband wireless communications, and wireless/wireline interworking. His recently coauthored text entitled *Wireless Communications and Networking* is currently under production by Prentice-Hall.

Dr. Mark is an IEEE fellow, a former editor of *IEEE Transactions on Communications*, currently a member of the Inter-Society Steering Committee of the IEEE/ACM Transactions on Networking, an editor of *Wireless Networks*, and an associate editor of *Telecommunication Systems*.

Shihua Zhu received his B.Sc. degree in radio techniques in 1982 from Xi'an Jiaotong University, China, his M.Sc. degree in telecommunication systems in 1984, and his Ph.D. degree in electric systems engineering in 1987, both from the University of Essex, United Kingdom. He joined Xi'an Jiaotong University in 1987 and is currently a professor and the dean of the School of Electronic and Information Engineering. In the summers of 1993, 1994, and 1995, he visited the Chinese University of Hong Kong where he involved in a wireless digital transceiver development. He also worked as a visiting professor at University of Waterloo between October 1998 and October 1999 where he carried out a research on resource management for wideband CDMA systems. His research interests are multiuser detection, nonlinear multipath channel modeling, resource management in mobile communications systems, and power control in wideband CDMA systems.

BIBLIOGRAPHY

- F. Adachi, M. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communications systems, *IEEE Commun. Mag.* 56–69 (Sept. 1998).
- E. Dahlman et al., WCDMA—The radio interface for future mobile multimedia communications, *IEEE Trans. Vehic. Technol.* 47(4): 1105-1118 (Nov. 1998).
- E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, UMTS/IMT-2000 based on wideband CDMA, *IEEE Commun.* Mag. 70-80 (Sept. 1998).
- E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* 48-54 (Sept. 1998).
- R. Gold, Optimal binary sequences for spread spectrum multiplexing, *IEEE Trans. Inform. Theory* **IT-13**: 619–621 (Oct. 1967).
- P. R. Larijani, J. W. Chinneck, and R. H. Hafez, Nonlinear power assignment in multimedia CDMA wireless networks, *IEEE Commun. Lett.* 2: 251-253 (Sept. 1998).
- J. W. Mark and S. Zhu, Power control and rate allocation in multirate wideband CDMA systems, *Proc. IEEE Wireless Communications and Networking Conf.*, 2000, pp. 168–172.

- T. Ojanpera and R. Prasad, An overview of air interface multiple access for IMT-2000/UMTS, *IEEE Commun. Mag.* 82–95 (Sept. 1998).
- R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications — a tutorial, *IEEE Trans. Commun.* COM-30: 855–884 (May 1982).
- D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudorandom and related sequences, *Proc. IEEE* 68: 598-619 (May 1980).
- R. D. Yates, A framework for uplink power control in cellular radio systems, *IEEE J. Select. Areas Commun.* 13: 1341–1347 (Sept. 1995).

WIRELESS AD HOC NETWORKS

Zygmunt J. Haas Jing Deng Ben Liang Panagiotis Papadimitratos S. Sajama Cornell University Ithaca, New York

1. INTRODUCTION

This section is reprinted from Ref. 1, pp. 221–225, © 2001 Addison Wesley Longman, Inc., by permission of Pearson Education, Inc.

1.1. The Notion of the Ad Hoc Networks

A mobile ad hoc network (MANET) is a network architecture that can be rapidly deployed without relying on preexisting fixed network infrastructure. The nodes in a MANET can dynamically join and leave the network, frequently, often without warning, and possibly without disruption to other nodes' communication. Finally, the nodes in the network can be highly mobile, thus rapidly changing the node constellation and the presence or absence of links. Examples of the use of the MANETs are

- *Tactical operation*—for fast establishment of military communication during the deployment of forces in unknown and hostile terrain
- *Rescue missions* for communication in areas without adequate wireless coverage
- *National security* for communication in times of national crisis, where the existing communication infrastructure is nonoperational due to a natural disaster or a global war
- Law enforcement—for rapid establishment of communication infrastructure during law enforcement operations
- *Commercial use*—for setting up communication in exhibitions, conferences, or sales presentations
- *Education* for operation of wall-free (virtual) class-rooms

• Sensor networks — for communication between intelligent sensors [e.g., microelectromechanical systems (MEMS)] mounted on mobile platforms

Nodes in the MANET exhibit nomadic behavior by freely migrating within some area, dynamically creating and tearing down associations with other nodes. Groups of nodes that have a common goal can create formations (clusters) and migrate together, similar to military units on missions or to guided tours on excursions. Nodes can communicate with each other at any time and without restrictions, except for connectivity limitations and subject to security provisions. Examples of network nodes are pedestrians, soldiers, or unmanned robots. Examples of mobile platforms on which the network nodes might reside are passenger cars, trucks, buses, tanks, trains, planes, helicopters, or ships.

MANETs are intended to provide a data network that is immediately deployable in arbitrary communication environments and is responsive to changes in network topology. Because ad hoc networks are intended to be deployable anywhere, existing infrastructure may not be present. The mobile nodes are thus likely to be the sole elements of the network. Differing mobility patterns and radio propagation conditions that vary with time and position can result in intermittent and sporadic connectivity between adjacent nodes. The result is a timevarying network topology.

MANETs are distinguished from other ad hoc networks by rapidly changing network topologies, influenced by the network size and node mobility. Such networks typically have a large span and contain hundreds to thousands of nodes. The MANET nodes exist on top of diverse platforms that exhibit quite different mobility patterns. Within a MANET, there can be significant variations in nodal speed (from stationary nodes to high-speed aircraft), direction of movement, acceleration/deceleration, or restrictions on paths (e.g., a car must drive on a road, but a tank does not). A pedestrian is restricted by built objects, while airborne platforms can exist anywhere in some range of altitudes. In spite of such volatility, the MANET is expected to deliver diverse traffic types, ranging from pure voice to integrated voice and image, and even possibly some limited video.

1.2. The Communication Environment and the MANET Model

The following are a number of assumptions about the communication parameters, the network architecture, and the network traffic in a MANET:

- Nodes are equipped with portable communication devices. Lightweight batteries may power these devices. Limited battery life can impose restrictions on the transmission range, communication activity (both transmitting and receiving) and computational power of these devices.
- Connectivity between nodes is *not* a transitive relation; if node *A* can communicate directly with node *B* and node *B* can communicate directly with node *C*, then node *A* may not, necessarily, be able to

communicate directly with node C. This leads to the hidden-terminal problem [2].

- A hierarchy in the network routing and mobility management procedures could improve network performance measures, such as the latency in locating a mobile. However, a physical hierarchy may lead to areas of congestion and is very vulnerable to frequent topological reconfigurations.
- We assume that nodes are identified by fixed IDs (e.g., based on IP [3] addresses).
- All the network nodes have equal capabilities. This means that all nodes are equipped with identical communication devices and are capable of performing functions from a common set of networking services. However, all nodes do not necessarily perform the same functions at the same time. In particular, nodes may be assigned specific functions in the network, and these roles may change over time.
- Although the network should allow communication between any two nodes, it is envisioned that a large portion of the traffic will be between geographically close nodes. This assumption is clearly justified in a hierarchical organization. For example, it is much more likely that communication will take place between two soldiers in the same unit, rather than between two soldiers in two different brigades.

A MANET is a *peer-to-peer* network that allows *direct* communication between any two nodes, when adequate radio propagation conditions exist between these two nodes and subject to transmission power limitations of the nodes. If there is no direct link between the source and the destination nodes, *multihop* routing is used. In multihop routing, a packet is forwarded from one node to another, until it reaches the destination. Of course, appropriate routing protocols are necessary to discover routes between the source and the destination, or even to determine the presence or absence of a path to the destination node. Because of the lack of central elements, distributed protocols have to be used.

The main challenges in the design and operation of the MANETs, compared to more traditional wireless networks, stem from the lack of a centralized entity, the potential for rapid node movement, and the fact that all communication is carried over the wireless medium. In standard cellular wireless networks, there are a number of centralized entities [e.g., the base stations, the mobile switching centers (MSCs), the home location register (HLR), and the visitor location register (VLR)]. In ad hoc networks, there is no preexisting infrastructure, and these centralized entities do not exist. The centralized entities in the cellular networks perform the function of coordination. The lack of these entities in the MANETs requires distributed algorithms to perform these functions. In particular, the traditional algorithms for mobility management, which rely on a centralized HLR/VLR, and the medium access control schemes, which rely on the base-station/MSC support, become inappropriate.

All communications between all network entities in ad hoc networks are carried over the wireless medium. Because the radio communications are vulnerable to propagation impairments, connectivity between network nodes is not guaranteed. In fact, intermittent and sporadic connectivity may be quite common. Additionally, as the wireless bandwidth is limited, its use should be minimized. Finally, as some of the mobile devices are expected to be handheld with limited power sources, the required transmission power should be minimized as well. Therefore, the transmission radius of each mobile (device) is limited, and channels assigned to mobiles are typically spatially reused. Consequently, since the transmission radius is much smaller than the network span, communication between two nodes often needs to be relayed through intermediate nodes; thus, multihop routing is used.

Because of the possibly rapid movement of the nodes and variable propagation conditions, network information, such as a route table, becomes obsolete quickly. Frequent network reconfiguration may trigger frequent exchanges of control information to reflect the current state of the network. However, the short lifetime of this information means that a large portion of this information may never be used. Thus, the bandwidth used for distribution of the routing update information is wasted. In spite of these attributes, the design of the MANETs still needs to allow for a high degree of reliability, survivability, availability, and manageability of the network.

On the basis of the discussion above, we require the following features for the MANETs:

- *Robust routing and mobility management algorithms* to increase the network's reliability and availability (e.g., to reduce the chances that any network component is isolated from the rest of the network)
- Adaptive algorithms and protocols to adjust to frequently changing radio propagation, network, and traffic conditions
- *Low-overhead algorithms and protocols* to preserve the radio communication resource
- Multiple (distinct) routes between a source and a destination to reduce congestion in the vicinity of certain nodes, and to increase reliability and survivability
- *Robust network architecture* to avoid susceptibility to network failures, congestion around certain nodes, and the penalty due to inefficient routing

In this article, we present a survey of techniques used to establish communications in MANETs. In particular, we concentrate on four areas: the medium access control (MAC) schemes, the routing protocols, the multicasting protocols, and the security schemes.

2. MAC-LAYER PROTOCOLS FOR AD HOC NETWORKS

Applicability of the existing MAC-layer protocol, in particular the family of the *carrier sense multiple access* (CSMA), to the radio environment is limited by



Figure 1. An example of the hidden-terminal problem.

the following two interference mechanisms: the hiddenterminal and the exposed-terminal problems.

The *hidden-terminal* problem occurs because the radio network, as opposed to other networks, such as a LAN, for instance, does not guarantee a high degree of connectivity. Thus, two nodes, which maintain connectivity to a third node, cannot, necessarily hear each other. Consider the situation in Fig. 1. Node α is in communication with node β . Node α is currently transmitting. Node γ wishes to communicate with node β as well. Following the CSMA protocol, node γ listens to the medium, but since there is an obstruction between node α and node γ , node γ does not detect node's α transmission, declaring the medium is free. Consequently, γ accesses the medium, causing collisions at β .

The second problem, the *exposed-terminal* problem, is depicted in Fig. 2. In the figure, node α is transmitting to node β , while node γ wants to transmit to node δ . Following the CSMA protocol, node γ listens to the medium, hears that node α transmits and defers from accessing the medium. However, there is no reason why node γ cannot transmit concurrently with the transmission of node α , as the transmission of node γ would not interfere with the reception at node β due to the distance between the two. The culprit here is, again, the fact that the collisions occur at the receiver, while the CSMA protocol checks the status of the medium at the transmitter.

In general, the hidden-terminal problem reduces the capacity of a network due to increasing the number of collisions, while the exposed-terminal problem reduces the network capacity due to the unnecessarily deferring nodes from transmitting.

Several attempts have been made in the literature to reduce the adverse effect of these two problems. The necessity of a dialog between the transmitting and the receiving nodes that preempts the actual transmission and that is referred to as the RTS/CTS dialog, has been



Figure 2. An example of the exposed-terminal problem.



Figure 3. The RTS/CTS dialog reduces the chances of collision.

generally accepted. The RTS/CTS dialog is depicted in Fig. 3. A node ready to transmit a packet, sends a short control packet, the *request to send* (RTS), with all nodes that hear the RTS defer from accessing the channel for the duration of the RTS/CTS dialog. The destination, on reception of the RTS, responds with another short control packet, the *clear to send* (CTS). All nodes that hear the CTS packet defer from accessing the channel for the duration of the DATA packet transmission. The reception of the CTS packet at the transmitting node acknowledges that the RTS/CTS dialog has been successful and the node starts the transmission of the actual data packet. Although the RTS/CTS dialog does not eliminate the hidden- and the exposed-terminal problems, it does provide some degree of improvement over the traditional CSMA schemes.

In what follows, we present a number of attempts to further improve the performance of the MAC-layer protocols for ad hoc networks.

2.1. The Multiple Access Collision Avoidance (MACA) Scheme

In multiple access collision avoidance (MACA), Karn [4] proposed the use of RTS/CTS dialog for collision avoidance on the shared channel. Through the use of the RTS/CTS dialog, the MACA scheme reduces the probability of data packet collisions caused by hidden terminals.

The function of the RTS packet in the MACA scheme is similar function to that of the packet preamble in the receiver initiated busy-tone multiple access (RI-BTMA) scheme [4.5]. The RI-BTMA scheme does not have the CTS packet, because it uses a "busy" tone to notify the communication initiator. Since the CTS packet may suffer from packet collisions, the notification from CTS packets is not as safe as that from the busy tone in the RI-BTMA scheme. An example is the reception failure of CTS packet at some hidden nodes because of transmissions from other nodes. These hidden nodes, without receiving any CTS packet notification, may transmit new RTS packets when the CTS packet sender is receiving its data packet. This leads to data packet collisions. It is clear that additional continuous notification is necessary to protect data packets.

2.2. The Multiple Access Collision Avoidance Wireless (MACAW) Scheme

Bharghavan [5] suggested the use of the RTS-CTS-DS-DATA-ACK message exchange for a data packet transmission in the MACAW protocol. Two new control packets were added to the packet train: DS and ACK packets. When the transmitter receives the CTS packet from its intended destination, it sends out a DS (data sending) packet before it transmits the data packet. The DS packet notifies neighbor nodes of the fact that a RTS/CTS dialog has been successful and a data packet will be sent. The ACK packet was implemented for immediate acknowledgment and the possibility of fast retransmission of collided data packets instead of upperlayer retransmission.

A new backoff algorithm, the multiple increase and linear decrease (MILD) algorithm, was also proposed in the paper to address the unfairness problem in accessing the shared channel. In the MILD backoff algorithm, successful nodes decrease their backoff interval by one step and unsuccessful nodes increase their backoff interval by multiplying them with 1.5. Backoff interval is also put into the header of the transmitted packet, so that the nodes overhearing successful packet transmission can copy the backoff interval on the packet into a local variable. Compared with the binary exponential backoff algorithm, the MILD algorithm has milder oscillation of the backoff intervals. Additional features of the MILD algorithm, such as multiple backoff intervals for different destinations, further improve the fairness performance of MACAW. The drawback of the MACAW scheme is inherited from the MACA scheme: the RTS/CTS packet collisions in a network with hidden terminals degrade its performance.

2.3. The Floor Acquisition Multiple Access (FAMA) Schemes

Fullmer and Garcia-Luna-Aceves [6] proposed the floor acquisition multiple access (FAMA) scheme. In FAMA, each ready node has to acquire the channel (the "floor") before it can use the channel to transmit its data packets. FAMA uses both carrier sensing and RTS/CTS dialog to ensure the acquisition of the "floor" and the successful transmission of the data packets. FAMA performs as well as MACA, when hidden terminals are present and as well as CSMA otherwise. In Ref. 7, FAMA was extended to FAMA-NPS (FAMA nonpersistent packet sensing) and FAMA-NCS (FAMA nonpersistent carrier sensing). FAMA-NPS requires nodes sensing packets to backoff. FAMA-NCS uses carrier sensing to keep neighbor nodes from transmitting while the channel is being used for data packet transmission. The length of the CTS packet is longer than that of the RTS packet, maintaining the dominance of CTS packets in the situation of collisions. Nodes can sense the carrier of the CTS packet when there is a collision between an RTS and a CTS packet and keep quiet; hence the data packet is protected at the receiver.

It was quantitatively shown [7] that FAMA-NPS did not perform well in situations with hidden terminals present, unless multiple transmissions of the CTS packet are used. The reason is the possible packet collisions resulting from hidden terminals. FAMA-NCS, by combining the carrier sensing and floor acquisition schemes together, outperforms nonpersistent CSMA and previous FAMA schemes in multihop networks.

2.4. The Dual-Busy-Tone Multiple Access (DBTMA) Scheme

In the DBTMA scheme [8], in addition to the use of an RTS packet, two out-of-band busy tones are used to notify neighbor nodes of the channel status. When a node is ready to transmit, it sets up its *transmit busy tone* and sends out an RTS packet to its intended receiver. On reception of the RTS packet, the receiver sets up a busy tone (the *receive busy tone*) and waits for the incoming data packet. The receive busy tone operates similarly to the busy tone of the RI-BTMA scheme. However, with the help of the second busy tone (the transmit busy tone), the probability of RTS packets colliding is decreased and the performance is improved.

The DBTMA scheme completely solved the hiddenterminal problems and the exposed-terminal problems. It forbids the hidden terminals to send any packet on the channel while the receiver is receiving the data packet. It allows the exposed terminals to initiate transmission by sending out the RTS packets. Furthermore, it allows the hidden terminals to reply to RTS packets by setting up the receive busy tone and initiate data packet reception.

3. ROUTING PROTOCOLS FOR AD HOC NETWORKS

Traditionally, the network routing protocols could be divided into proactive protocols and reactive protocols. Proactive protocols continuously learn the topology of the network by exchanging topological information among the network nodes. Thus, when there is a need for a route to a destination, such route information is available immediately. The early protocols that were proposed for routing in ad hoc networks were proactive distance vector protocols based on the Distributed Bellman-Ford (DBF) algorithm [9]. To address the problems of the DBF algorithm-convergence and excessive control traffic, which are especially an issue in resource-poor ad hoc networks-modifications were considered [10-12]. Yet another approach taken to address the convergence problem is the application of the link-state protocols to the ad hoc environment. An example of the latter is the optimized link-state routing protocol (OLSR) [13]. Yet another approach taken by some researchers is the proactive path-finding algorithms. In this approach, which combines the features of the distance vector and linkstate approaches, every node in the network constructs a minimum spanning tree (MST), using the information of the MSTs of its neighbors, together with the cost of the link to its neighbors. The path-finding algorithms allow to reduce the amount of control traffic, to reduce the possibility of temporary routing loops, and to avoid the "counting-to-infinity" problem. An example of this type of routing protocols is the wireless routing protocol (WRP) [14,15].

The main issue with the application of proactive protocols to the ad hoc networking environment stems from the fact that as the topology continuously changes, the cost of updating the topological information may be prohibitively high. Moreover, if the network activity is low, the information about the actual topology may even not be used and the investment of limited transmission and computing resources in maintaining the topology is lost.

On the other end of the spectrum are the *reactive* routing protocols, which are based on some type of "query-reply" dialog. Reactive protocols do not attempt to continuously maintain the up-to-date topology of the network. Rather, when the need arises, a reactive protocol invokes a procedure to find a route to the destination; such a procedure involves some sort of flooding the network with the route query. As such, such protocols are often also referred to as on-demand. Examples of reactive protocols include the temporally ordered routing algorithm (TORA) [16], the dynamic source routing (DSR) [17], and ad hoc on-demand distance vector (AODV) [18]. In TORA, the route replies use controlled flooding to distribute the routing information through a form of a *directed acyclic* graph (DAG), which is rooted at the destination. The DSR and the AODV protocols, on the other hand, use unicast to route the reply back to the source of the routing query, along the reverse path of the query packet. The reversed path is "inscribed" into the query packet as "accumulated" route in the DSR and is used for source routing. In AODV, the path information is stored as the "next hop" information within the nodes on the path. Although the reactive approach can lead to less control traffic, as compared with proactive distance vector or linkstate schemes, in particular, when the network activity is low and the topological changes frequent, the amount of traffic can still be significant at times. Moreover, due to the networkwide flooding, the delay associated with reactive route discovery may be considerable as well.

Thus, both of the routing "extremes," the proactive and the reactive schemes, may not perform best in a highly dynamic networking environment, such as in ad hoc networks. Although proactive protocols can produce the required route immediately, they may waste too much of the network resources in the attempt to always maintain the updated network topology. The reactive protocol, on the other hand, may reduce the amount of used network resources, but may encounter excessive delay in the flooding of the network with routing queries. Another approach to address the routing problem is through the hybrid protocols, which incorporate some aspects of the proactive and some aspects of the reactive protocols. The zone routing protocol (ZRP) [19] is an example of the hybrid approach. In ZRP, each node proactively maintains the topology of its close neighborhood only, thus reducing the amount of control traffic relative to the proactive approach. To discover routes outside its neighborhood, the node reactively invokes a generalized form of controlled flooding, which reduces the route discovery delay, as compared with purely reactive schemes. The size of the neighborhood is a single parameter that allows optimizing the behavior of the protocol based on the degree of nodal mobility and the degree of network activity.

In what follows, we present a number of examples of routing protocols that were developed for the ad hoc networking environment.

3.1. Single-Scope Routing Protocols

3.1.1. Advantages and Disadvantages. The main advantage of the single-scope routing protocols, in comparison

with the multiscope routing protocols, is their lower complexity. There is no distinction of nearby or faraway nodes, and there is no need to maintain a hierarchical structure. Therefore, they are generally simpler to implement, both in simulations and in practical systems. The current activities within the IETF MANET group predominantly involve single-scope routing.

However, inefficient resource management can result from treating nodes equally, regardless of their relative location. For example, it may not be necessary for a node to maintain very accurate link-state tables or route caching information of faraway nodes. Therefore, the single-scope routing protocols may not scale well as the network size increases.

The single-scope routing protocols can be categorized into reactive (or on-demand) and proactive (or tabledriven) ones. The main advantage of the reactive protocols is that no routing-table updating is required, unless a route is used. Therefore, battery power and wireless bandwidth can be conservatively utilized. However, when a route is needed, the source node needs to query for the route. That can lead to routing delay. Furthermore, an efficient route querying mechanism is required in order to prevent overloading the network with query packets.

On the other hand, the proactive protocols generally provide a source node with readily available routes to all other nodes. They incur no routing delay or query traffic. The disadvantage of the proactive protocols is that they may incur unnecessary control traffic in maintaining upto-date topology information, whether that information is needed for routing or not.

3.1.2. Reactive/On-Demand Routing Protocols

3.1.2.1. Ad Hoc On-Demand Distance Vector Routing (AODV). AODV [19] incorporates the destination sequence number technique of destination-sequenced distance vector (DSDV) routing into an on-demand protocol. (DSDV is discussed in the sequel.)

Each node keeps a next-hop routing table containing the destinations to which it currently has a route. A route expires, if it is not used or reactivated for a threshold amount of time.

If a source has no route to a destination, it broadcasts a route request (RREQ) packet using an expanding ring search procedure, starting from a small time-to-live value (maximum hop count) for the RREQ, and increasing it if the destination is not found. The RREQ contains the last seen sequence number of the destination, as well as the source node's current sequence number. Any node that receives the RREQ updates its next-hop table entries with respect to the source node. A node that has a route to the destination with a higher-sequence number than the one specified in the RREQ unicasts a route reply (RREP) packet back to the source. Upon receiving the RREP packet, each intermediate node along the RREP routes updates its next-hop table entries with respect to the destination node, dropping the redundant RREP packets and those RREP packets with a lower destination sequence number than one seen previously.

When an intermediate node discovers a broken link in an active route, it broadcasts a route error (RERR) packet to its neighbors, which in turn propagate the RERR packet upstream toward all the nodes that have an active route using the broken link. The affected source can then reinitiate route discovery, if the route is still needed.

3.1.2.2. Dynamic Source Routing (DSR). DSR [17] is a source routing on-demand protocol with various efficiency improvements. In DSR, each node keeps a route cache that contains full paths to known destinations. If a source has no route to a destination, it broadcasts a route request packet to its neighbors. Any node receiving the route request packet and without a route to the destination appends its own ID to the packet and rebroadcasts the packet. If a node receiving the route request packet has a route to the destination, the node replies to the source with a concatenation of the path from the source to itself and the path from itself to the destination. If the node already has a route to the source, the route reply packet will be sent over that route. Otherwise, depending on the underlining assumption of the directionality of links, the route reply packet can be sent over the reversed sourceto-node path, or piggybacked in the node's route request packet for the source.

When an intermediate node discovers a broken link in an active route, it sends a route error packet to the source, which may reinitiate route discovery if an alternate route is not available.

DSR has efficiency improving features. One such feature is the *promiscuous* mode, in which a node listens to route request, reply, or error messages not intended to itself and updates its route cache correspondingly. Another DSR feature is the *expanding ring search* procedure, in which the route request packets are sent with a maximum hop count, which can be increased if the destination is not found within the hop-count limit. Finally, adding *jitter* in sending the route reply messages to prevent *route reply storms* and *packet salvaging* to extract correct routes from route error packets, are yet two other features that improve DSR performance.

3.1.2.3. Temporally Ordered Routing Algorithm (TORA). TORA [20] is a merger of the proactive link-reversal algorithm for destination-oriented directional acyclic graph creation [21] and the on-demand query-reply mechanism of lightweight mobile routing (LMR) [22].

In TORA, routes to a destination are defined by a directional acyclic graph (DAG) rooted at the destination. Each link in the network is assumed to be bi-directional, but in order to form the DAG with respect to a destination, a logical direction of the link is defined by giving *height* values to the two nodes at the ends of the link. Since time is part of the height value, TORA requires synchronized clocks across all nodes.

If a source has no route to a destination (i.e., the source node has no outgoing edge in the DAG), it broadcasts a route query packet (QRY), which is propagated outward by its neighbors. After receiving the QRY, a node that has a route to the destination broadcasts a route update packet (UPD) containing its own height. Receiving the UPD, each node that doesn't have a route to the destination updates its height to reflect the creation of an outgoing edge. Route maintenance is achieved through height adjustment and UPD exchange. Network partition can be detected by a node receiving UPDs reflected from the partition boundary, in which case a clear message (CLR) is used to update all routes within the partition.

TORA also supports a proactive mode, in which the destination initiates the route creation process by sending a packet that is processed and forwarded by the neighboring nodes.

3.1.3. Proactive/Table-Driven

3.1.3.1. Destination-Sequenced Distance-Vector Routing (DSDV). DSDV [12] provides improvements over the conventional Bellman–Ford distance vector protocol. It eliminates route looping, increases convergence speed, and reduces control message overhead.

In DSDV, each node maintains a next-hop table, which it exchanges with its neighbors. There are two types of next-hop table exchanges: periodic full-table broadcast and event-driven incremental updating. The relative frequency of the full-table broadcast and the incremental updating is determined by the node mobility.

In each data packet sent during a next-hop table broadcast or incremental updating, the source node appends a sequence number. This sequence number is propagated by all nodes receiving the corresponding distance vector updates, and is stored in the next-hop table entry of these nodes. A node, after receiving a new next-hop table from its neighbor, updates its route to a destination only if the new sequence number is larger than the recorded one, or if the new sequence number is the same as the recorded one, but the new route is shorter.

In order to further reduce the control message overhead, a *settling time* is estimated for each route. A node updates its neighbors with a new route only if the settling time of the route has expired and the route remains optimal.

3.1.3.2. Wireless Routing Protocol (WRP). This protocol [15] provides improvements over the Bellman–Ford distance vector protocol. It reduces the amount of route looping, and has a mechanism to ensure the reliable exchange of update messages.

In WRP, each node maintains a distance table matrix, which contains all destination nodes, and, for each destination node, all neighbors through which the destination node can be reached. For each neighbor-destination pair, if a route exists, the route length is recorded. Also recorded is the *predecessor*, the last node along a route before the destination node.

Each node's neighbor broadcasts its current best route to selected destinations on an event-driven incremental basis. After a broadcast, acknowledgments are expected from all neighbor nodes. If some acknowledgments are missing, the broadcast will be repeated, with a *message retransmission list* specifying the subset of neighbors that need to respond. A node, after receiving the route updating packets from a neighbor, updates its own routing table only if the consistency of the new information is checked against the predecessor information from all its neighbors.

3.2. Multiscope Routing Protocols

3.2.1. Advantages and Disadvantages. The multiscope routing protocols distinguish nodes by their relative positions. More resource is devoted to maintaining the topology information of more nearby, and hence more frequently used, parts of the network. Therefore, scalability is the main advantage of the multiscope routing protocols.

Their disadvantage is their relative complexity in comparison with the single-scope routing protocols. Ranking mechanisms that distinguish the nodes are required. Furthermore, they generally need to be reconfigurable, in order to adapt to the changing network topology and the varying node traffic and movement patterns.

Multiscope routing can be categorized into the flat protocols and the hierarchical protocols. The main advantage of the flat protocols, in comparison with the hierarchical ones, is that they do not require specialized nodes. All nodes serve the same set of functions. Therefore, they are relatively simple to implement, and they avoid the control message overhead and nonuniform loading involved in node specialization. However, since the flat structure does not have special nodes that can provide locally centralized functionality, the nodes between nearby local scope exchange link information in a strictly distributive manner. Thus, the lack of coordination can lead to inefficiency.

On the other hand, the hierarchical protocols utilize specialized nodes, such as the cluster heads, group leaders, or the route gateways, to coordinate the dissemination of local link information. Furthermore, the relative position of the specialized nodes can provide directional guidance to routing between the regular nodes. However, the dynamic maintenance of the hierarchy can potentially consume a large amount of the battery power and wireless bandwidth from routing itself, especially when the network is highly mobile. Furthermore, mechanisms are needed to avoid overloading the local controllers and to alleviate the traffic hot spots.

3.2.2. Flat Routing Protocols

3.2.2.1. Zone Routing Protocol (ZRP). This protocol [18] provides a hybrid routing framework that is locally proactive and globally reactive. Each node proactively advertises its link state within a fixed number of hops, called the *zone radius*. These local advertisements give each node an updated view of its routing zone — the collection of all nodes and links that are reachable within the zone radius. The routing zone nodes that are at the minimum distance of the zone radius are called *peripheral nodes*. The peripheral nodes represent the boundary of the routing zone and play an important role in zone-based route discovery. Each node has an associated routing zone, and routing zones of neighboring nodes overlap.

ZRP uses the knowledge of the routing zone connectivity to guide its global route discovery. Rather than blindly broadcasting route queries from a node to all its neighbors, ZRP employs a service called *bordercasting*, which directs the route request from a node to its peripheral nodes via multicast. Special query control mechanisms are used to identify those peripheral nodes that have been covered by the route query (i.e., that belong to the routing zone of a node that already has bordercast the query) and prune them from the bordercast's query distribution tree. This encourages the query to propagate outward, away from its source and away from covered regions of the network.

Routing zones also help improve the quality and survivability of discovered routes, by making them more robust to changes in network topology. Once routes have been discovered, routing zones offer enhanced, real-time, route maintenance. Multiple hop paths within the routing zone can bypass link failures. Similarly, sub optimal route segments can be identified and traffic can be rerouted along shorter paths.

3.2.2.2. Optimized Link State Routing (OLSR). OLSR [23] is a link-state protocol, where the link information is disseminated through an efficient flooding technique.

The key concept in OLSR is *multipoint relay* (MPR). A node's MPR set is a subset of its neighbors, whose combined radio range covers all nodes two hops away. Heuristics are proposed for each node to determine its minimum MPR set based on its two-hop topology. Each node obtains the two-hop topology through its neighbors' periodic broadcasting of "Hello" packets containing the neighbors' lists of neighbors.

As with a conventional link-state protocol, a node's link information update is propagated throughout the network. However, in OlSR, when a node forwards a link updating packet, only those neighbors in the node's MPR set participate in forwarding the packet (similar to ZRP's border-casting with *one-hop* zone radius).

Furthermore, a node only originates link updates concerning those links between itself and the nodes in its MPR set. Therefore, routes are computed using a node's partial view of the network topology.

3.2.2.3. Fisheye State Routing (FSR). The fisheye routing concept is based on the premise that changes in a network region's topology have less effect on a router's packet forwarding decisions as the distance (in hops) between the router and the network increases. This relationship can be exploited in order to reduce routing traffic by relaying topology updates for distant regions less often than updates for nearby regions. Given an approximate view of the distant parts of the network, a node can forward a packet in the proper direction toward the destination. As the packet progresses toward the destination, the view of the destination's region becomes more accurate, providing for more precise packet forwarding.

This fisheye technique is applied in the fisheye state routing (FSR) protocol [24], an adaptation of the global state routing (GSR) [25]. In the original GSR protocol, link-state information is propagated through the network by periodic link state table exchanges between neighbors. In FSR, a node exchanges individual link state table entries at different rates, depending on the distance to the link's source. In particular, FSR defines *scopes* of increasing radii (in hops) around each node. A node relays a link state table entry if the link's source lies within the largest scope covered by the current table exchange. The first level (innermost) scope is covered by every table exchange. The *k*th-level scope is covered by every X_k th interval, where X_k is an integer multiple of X_{k-1} . This relationship ensures that an exchange covering a level *k*th scope coincides with the more frequent updates of all the interior scopes.

3.2.3. Hierarchical Routing Protocols

3.2.3.1. Core-Extraction Distributed Ad Hoc Routing (CEDAR). CEDAR [26] employs a set of core nodes, at least one of which is within one hop of each node, in its routing mechanism. The core nodes are selected using a highest-degree scheme. A core node dominates each noncore node. Through periodic updating, the noncore nodes maintain a list of the IDs of their neighbors and their neighbors' respective dominators. The state information of each link is disseminated toward the core nodes away from the link, and the higher is the capacity of link, the further the information travels. Each core node keeps a local link-state table containing only the stable, high capacity, and nearby links.

Global route search is carried out reactively. Similar to CBRP, the dominator of a source node determines a *core path* to the dominator of the destination node by an efficient flooding over the core. Using its local link state, the dominator of the source computes a "shortest-widest-furthest" QoS-admissible path to an intermediate node, along the core path toward the destination. It then sends a route forwarding request to the dominator of the intermediate node, which then starts the same QoS-admissible path search using its own local link-state table. The process continues until the QoSadmissible path reaches the destination. Source routing is then carried out to forward the data packets.

3.2.3.2. Zone-Based Hierarchical Link State (ZHLS). In ZHLS [27], the system coverage area is divided into nonoverlapping physical zones. The nodes are equipped with geographic location devices such as the GPS receivers, so that each node can determine its zone membership by comparing its physical location with the zone map. Furthermore, if the nodes within a zone are partitioned, logical subzones are created, each containing one of the partitions. Every node maintains an intrazone routing table and an interzone routing table. The intrazone routing table enables a node to reach all the other nodes within the zone. Interzone communications are carried out through the gateway nodes near the zone edges. The gateway nodes broadcast the status of the virtual links between zones to the entire network. A node aggregates all gateway broadcasts to form the interzone routing table.

When a source node needs to transmit data to a destination node outside the source node's zone, global zone query is used to determine the zone identity of the destination node. Using its interzone routing table, the source node sends the query to all zones in the network. After receiving the query message, the gateway nodes, whose zone contains the destination node sends back to the source node the destination node's zone identity. The source node then sends out the data packets with the destination node's zone ID and node ID specified in their headers. The packets are then forwarded to the destination node according to both the interzone and the intrazone routing tables at the intermediate nodes.

3.2.3.3. Landmark Ad Hoc Routing (LANMAR). The original landmark scheme for wired networks was proposed by Tsuchiya [28]. LANMAR [29] adopts that scheme for ad hoc network routing. In LANMAR, the network consists of predefined logical subnets, each with a preselected *landmark*. All nodes in a subnet are assumed to move as a group, and they remain connected to each other via *fisheye state routing*.

The routes to the landmarks, and hence the corresponding subnets, are proactively maintained by all nodes in the network through the exchange of distance vectors. Every node has a lifetime hierarchical address, identifying the subnet where it belongs. A source node specifies the hierarchical address of a destination node in the data packet headers. The packets are then forwarded toward the landmark of the subnet, where the destination node belongs. When a packet reaches a node in the subnet, where the destination node belongs, the node forwards the packet to the destination node using its subnet routing table.

3.3. Geographically Routed Protocols

3.3.1. Location-Aided Routing (LAR). In LAR [30], a source node estimates the range of a destination's location, based on the destination's last reported velocity, and broadcasts route request only to nodes within a geographically defined *request zone*. LAR requires each node to obtain its geographic location through external devices such as GPS.

3.3.2. Distance Routing Effect Algorithm for Mobility (DREAM). In DREAM [31], each node obtains its geographic location through external devices such as GPS, and periodically transmits its location coordinates to other nodes in the network. The period of location transmission depends on the node's velocity and the geographic distance to nodes to which the location information is intended.

A source sends a data packet to a subset of its neighbors in the direction of the destination. The intermediate nodes similarly forward the data packet towards the destination.

4. MULTICASTING PROTOCOLS FOR AD HOC NETWORKS

Multicasting is an efficient communication tool for use in multipoint applications. Many of the proposed multicast routing protocols, both for the Internet and for ad hoc networks, construct trees over which information is transmitted. Using trees is evidently more efficient than brute-force approach of sending the same information from the source individually to each receiver. Another benefit of using trees is that routing decisions at the intermediate nodes become very simple; a router in a multicast tree that receives a multicast packet over an in-tree interface forward the packet over the rest of its in-tree interfaces.

- Shortest-path tree algorithms [9]
- *Minimum-cost tree* algorithms [33,34]
- Constrained tree algorithms [35,36]

There are two fundamental approaches in designing multicast routing—one is to minimize the distance (or cost) from the sender to each receiver individually (shortest path tree algorithms), and the other is to minimize the overall (total) cost of the multicast tree. Practical considerations lead to a third category of algorithms, which try to optimize both constraints using some metric (minimum cost trees with constrained delays). The majority of multicast routing protocols in the Internet is based on shortest-path trees, because of their ease of implementation. Also, they provide minimum delay from sender to receiver, which is desirable for most real-life multicast applications. However shared trees are used in some more recent protocols (e.g., PIM [37] and CBT [38]), in order to minimize the state stored in the routers.

Multicasting in ad hoc networks is more challenging than in the Internet, because of the need to optimize the use of several resources simultaneously: (1) nodes in ad hoc networks are battery-power-limited and data travel over the air where wireless resources are scarce; (2) there is no centralized access point or existing infrastructure (as in the cellular network) to keep track of the node mobility; and (3) the status of communication links between nodes is a function of their positions, transmission power levels, and so on. The mobility of routers and randomness of other connectivity factors lead to a network with a potentially unpredictable and rapidly changing topology. This means that by the time a reasonable amount of information about the topology of the network is collected and a tree is computed, there may be very little time before the computed tree becomes useless.

Work on multicast routing in ad hoc networks gained momentum in the mid-1990s. Some early approaches to provide multicast support in ad hoc networks consisted of adapting the existing Internet multicasting protocols; for example, Shared Tree Wireless network Multicast [39]. On the other hand, ODMRP [40], AMRIS [41], CAMP [42], and others [43-51] have been designed specifically for ad hoc networks. ODMRP is a mesh-based, on-demand protocol that uses a soft state approach for maintenance of the message transmission structure. It exploits the robustness of mesh structure to frequent route failure and gains stability at the expense of bandwidth. The coreassisted mesh protocol (CAMP) attempts to remedy this excessive overhead, while still using a mesh by using a core for route discovery. AMRIS constructs a shared delivery tree rooted at a node, with ID numbers increasing as they radiate from the source. Local route recovery is made possible due to this property of ID numbers, hence reducing the route recovery time and also confining route recovery traffic to the region of link failures.

In what follows, we present a number of examples of multicasting protocols that were developed for the ad hoc networking environment.

4.1. Core Assisted Mesh Protocol (CAMP)

CAMP [42] builds and maintains a multicast mesh for information distribution within each multicast group. A router is allowed to accept unique packets coming from any neighbor in the process of forwarding packets through the mesh. Because a member router of a mesh has redundant paths to any other router in the mesh, this protocol is more resilient to topology changes than tree based protocols.

Cores are used to limit the control traffic needed for receivers to join multicast groups. In contrast to CBT [38], one or multiple cores can be defined for each mesh and cores need not be part of the mesh of their group. Routers can join a group even if all associated cores become unreachable using an expanded ring search. CAMP ensures that all reverse shortest paths between sources and receivers are part of groups mesh by means of "heartbeat" messages. In the event of link failure and partition, the operation of mesh components continues. Different components merge by sending join requests to cores as soon as connectivity with a core is reestablished.

CAMP is designed to support very dynamic ad hoc networks. According to the performance analysis presented in [42], this article, CAMP performs better than the on-demand multicast routing protocol (ODMRP) in terms of percentage of packets lost by routers, average packet delay, and total number of control packets received by each router. (ODMRP is described below.)

4.2. Multicast Operation of Ad Hoc On-Demand Distance Vector Routing Protocol

This is an extension of AODV to support multicasting and it builds multicast trees on demand to connect group members. Route discovery in MAODV [43] follows a route request/route reply discovery cycle. As nodes join the group, a multicast tree composed of group members is created. Multicast group membership is dynamic and group members are routers in the multicast tree. Link breakage is repaired by downstream node broadcasting a route request message. The control of a multicast tree is distributed, so there is no single point of failure. One big advantage claimed is that since AODV offers both unicast and multicast communication; route information, when searching for a multicast route can also increase unicast routing knowledge and vice versa.

In [43], an ad hoc network consists of laptops in a room (50-100 m wide, 10-m range) talking to each other, moving at a rate of 1 m/s. The results presented only verify working of AODV and do not compare performance with other multicasting protocols. [43] shows that AODV attains a high output ratio and is able to offer this communication with a minimum of control packet overhead. It also demonstrates its operation under frequent network partitions.

4.3. AMRIS: A Multicast Protocol for Ad Hoc Wireless Networks

AMRIS [41] is an on-demand protocol that constructs a shared delivery tree to support multiple senders and receivers within a multicast session. Each participant in the multicast session has a session-specific multicast session member id (*msm-id*). These msm-ids increase in numerical value as they radiate away from a central node known as SID. Tree initialization is done by the SID broadcasting a new session message. All nodes of the network calculate their msm-id to be larger than the msmid of the node they received the new session message from. There are beacon messages exchanged between nodes, which help a node to calculate its new msm-id after it moves to a new location.

AMRIS does not depend on the unicast routing protocol to provide routing information to other nodes, since it maintains a neighbor-status table. It is the child's responsibility to reconnect to the tree if a link failure occurs. If it has potential parents, that is, neighboring nodes with lower msm-ids, it sends a join request to them, which in turn try to join the tree in the same way. If there are no potential parents, the node transmits a join request message.

In the simulation given by Wu and Tay [41], the authors vary membership from 50 to 100 and the speed of nodes is up to 20 m/s. The simulation results presented study various performance parameters in terms of network conditions. The paper studied the effect of beacon intervals, membership sizes and mobility on packet delivery ratio and concludes that there is an optimum beacon interval. Control overhead is verified to be higher, when the beacon interval is small. They also show that the relationship between end-to-end packet delay and packet delivery ratio is robust with respect to membership.

4.4. AMRoute: Ad Hoc Multicast Routing Protocol

AMRoute [44] presents an approach for robust IP multicast in ad hoc networks by exploiting user-multicast trees and dynamic logical cores. It creates a bidirectional shared tree for data distribution using only group senders and receivers as tree nodes. Unicast tunnels are used as tree links to connect neighbors in the user-multicast tree. Hence the AMRoute protocol does not need to be supported by other nodes in the network. Also the tree structure does not change even in the case of dynamic network topology and hence reduces signaling. Each node is aware of its tree neighbors only and forwards data on the tree links to its neighbors. This saves node resources.

Certain tree nodes are designated by AMRoute as logical cores and are responsible for initiating and managing the signaling component of AMRoute, such as detection of group members and tree setup. Unlike CBT and PIM-SM, they are not a central point for data distribution and can migrate dynamically among member nodes. Hence there is no single point of failure. Like DVMRP, AMRoute provides robustness by periodic flooding for tree construction. However, AMRoute periodically floods a small signaling message instead of data.

AMRoute simulations were done with TORA as the underlying unicast protocol. The mobility of the network was emulated by keeping the node location fixed and breaking/connecting links between neighbors. The simulation results show that broadcasting signaling traffic generated by AMRoute is independent of the group size and inversely proportional to the network mobility. Unicast signaling traffic is proportional to the group size and inversely proportional to the network mobility. Total signaling traffic is independent of the data rate. Both signaling traffic and join latency are relatively low for typical group sizes. They verify that group members receive a high proportion of data multicast by a sender, even in the case of a dynamic network.

4.5. ODMRP: On-Demand Multicast Routing Protocol

ODMRP [40] is a mesh based, rather than a conventional tree based, scheme and uses a forwarding group concept (only a subset of nodes forwards the multicast packets via scoped flooding). By maintaining a mesh instead of a tree, the drawbacks of multicast trees in ad hoc networks, like frequent tree reconfiguration and non-shortest path in a shared tree, are avoided. ODMRP applies on-demand routing techniques to avoid channel overhead and to improve scalability.

The source starts a session by flooding a "join data" control packet with data payload attached, which is subsequently broadcast at regular intervals to the entire network to refresh membership information and update the routes. The mesh is created by the replies of receivers to this packet received via various paths. When receiving a multicast data packet, a node forwards it only when it is not a duplicate, hence minimizing traffic overhead.

Because the nodes maintain soft state, finding the optimal flooding interval is critical to ODMRP performance. ODMRP uses location and movement information to predict the duration of time that routes will remain valid. With the predicted time of route disconnection, a "join data" packet is flooded, when route breaks of ongoing data sessions are imminent. Lee et al. [40] compare DVMRP with ODMRP and show that the latter is better suited for ad hoc networks in terms of bandwidth utilization.

4.6. MCEDAR: Multicasting Core-Extraction Distributed Ad Hoc Routing

This scheme [50] is a multicast extension of CEDAR, which was a routing scheme proposed for unicast communication in ad hoc networks. MCEDAR relies on the core extraction and the core broadcast components of the CEDAR architecture. The core-extraction algorithm used is a distributed heuristic for finding a good approximation to a minimum dominating set. Each core node has the following state stored in it: its nearby core nodes and the nodes it dominates (i.e., each core node has enough local information to reach the domain of its nearby nodes and set up virtual links). Core broadcast is used instead of flooding in order to discover a route to the destination. This is done by making each node cache every RTS and CTS packet that it hears on the channel for core broadcast packets. So a node knows whether a packet has been received by the destination already. If it has, it does not transmit the packet and hence suppresses the duplicate transmission.

The infrastructure for a multicast group resides entirely within the core broadcast mechanism, which is used to perform data forwarding. Each multicast group extracts a subgraph of the core graph to function as "mgraph." Data forwarding is done on the mgraph using the core broadcast mechanism. In this way the forwarding is tree-based, although the structure is robust, because it is a mesh.

5. SECURITY OF AD HOC NETWORKS

The provision of security services in the MANET context faces a set of challenges specific to this new technology. The insecurity of the wireless links, energy constraints, relatively poor physical protection of nodes in a hostile environment, and the vulnerability of statically configured security schemes have been identified in the literature [52,53] as such challenges. Nevertheless, the single most important feature that differentiates MANET is the absence of a fixed infrastructure. No part of the network is dedicated to support individually any specific network functionality; routing (topology discovery, data forwarding) is the most prominent example. Additional examples of functions that cannot rely on a central service, and that are also of high relevance to this work, are naming services, certification authorities (CAs), directory, and other administrative services.

Even if such services were assumed, their availability would not be guaranteed, due either to the dynamically changing topology that could easily result in a partitioned network, or to congested links close to the node acting as a server. Furthermore, performance issues, such as delay constraints on acquiring responses from the assumed infrastructure, would pose an additional challenge.

The absence of infrastructure and the consequent absence of authorization facilities impede the usual practice of establishing a line of defense, separating nodes into trusted and nontrusted. Such a distinction would have been based on a security policy, the possession of the necessary credentials and the ability for nodes to validate them. In the MANET context, there may be no ground for an a priori classification, since all nodes are required to cooperate in supporting the network operation, while no prior security association can be assumed for all the network nodes. Additionally, in MANET, freely roaming nodes form transient associations with their neighbors; join and leave MANET sub-domains independently and without notice. Thus it may be difficult in most cases to have a clear picture of the ad hoc network membership. Consequently, especially in the case of a large-size network, no form of established trust relationships among the majority of nodes could be assumed.

In such an environment, there is no guarantee that a path between two nodes would be free of malicious nodes, which would not comply with the employed protocol and attempt to harm the network operation. The mechanisms currently incorporated in MANET routing protocols cannot cope with disruptions due to malicious behavior. For example, any node could claim that it is one hop away from the sought destination, causing all routes to the destination to pass through itself. Alternatively, a malicious node could corrupt any in-transit route request (reply) packet and cause data to be misrouted.

The presence of even a small number of adversarial nodes could result in repeatedly compromised routes, and, as a result, the network nodes would have to rely on cycles of timeout and new route discoveries to communicate. This would incur arbitrary delays before the establishment of a noncorrupted path, while successive broadcasts of route requests would impose excessive transmission overhead. In particular, intentionally falsified routing messages would result in a denial-of-service (DoS) experienced by the end nodes.

Despite the fact that security of MANET routing protocols is envisioned to be a major roadblock in commercial application of this technology, only a limited number of works has been published in this area. Below, we review some schemes related to the problem of incorporating security provisions within the context of ad hoc communication.

5.1. Overview of Security Schemes for Ad Hoc Networks

Efforts to incorporate security measures in the ad hoc networking environment have concentrated mostly on the aspect of data forwarding, disregarding the aspect of topology discovery. On the other hand, solutions that target route discovery have been based on approaches for fixed-infrastructure networks, defying the particular MANET challenges.

For the problem of secure data forwarding, two mechanisms that (1) detect *misbehaving* nodes and report such events and (2) maintain a set of metrics reflecting the past behavior of other nodes [54] have been proposed to alleviate the detrimental effects of packet dropping. Each node may choose the "best" route, composed of relatively well behaved nodes, namely, nodes that do not have history of avoiding forwarding packets along established routes. Among the assumptions of the abovementioned work [54] are a shared medium, bidirectional links, use of source routing (i.e., packets carry the entire route that becomes known to all intermediate nodes), and no colluding malicious nodes. Nodes operating in promiscuous mode overhear the transmissions of their successors and may verify whether the packet was forwarded to the downstream node and check the integrity of the forwarded packet. On detection of a misbehaving node, a report is generated and nodes update the rating of the reported misbehaving node. The ratings of nodes along a well-behaved route are periodically incremented, while reception of a misbehavior alert dramatically decreases the node rating.¹ When a new route is required, the source node calculates a path metric equal to the average of the ratings of the nodes in each of the route replies and selects the route with the highest metric.

The detection mechanism exploits two features that frequently appear in MANET: the use of a shared channel and source routing. Nevertheless, the plausibility of this solution could be questioned for several reasons, and, indeed, the authors provide a short list of scenarios of incorrect detection. The possibility of falsely detecting misbehaving nodes could easily create a situation with

¹The initial rating, 0.5, is increased by 0.01 every 200 ms. Suspected nodes have a rating equal to -100, with the option for a long timeout period after which the negative rating is changed back to a positive value.

many nodes falsely suspected for a long period of time. In addition, the metric construction may lead to a route choice that includes a suspected node, if, for example, the number of hops is relatively high, so that a low rating is "averaged out." Finally, the most important vulnerability is the proposed feedback itself; there is no way for the source, or any other node that receives a misbehavior report to validate its authenticity or correctness. Consequently, the simplest attack would be to generate fake alerts and eventually disable the network operation altogether. The protocol attempts new route discoveries, when none of the route replies is free of suspected nodes, with the excessive route request traffic degrading the network performance. At the same time, the adversary can falsely accuse a significant fraction of nodes within the timeout period related to reinstating from a negative rating and, essentially, partition the network.

A different approach [55] is to provide incentive to nodes, so that they comply with protocol rules to properly relay user data. The concept of fictitious currency is introduced, in order to endogenize the behavior of the assumed greedy nodes, which would forward packets in exchange for currency. Each intermediate node purchases from its predecessor the received data packet and sells it to its successor along the path to the destination. Eventually the destination pays for the received packet.² This scheme assumes the existence of an overlaid geographic routing infrastructure and a public key infrastructure (PKI). All nodes are preloaded with an amount of currency, have unique identifiers, are associated with a pair of private/public keys, and all cryptographic operations related to the currency transfers are performed by a physically tamper-resistant module. The applicability of the scheme, which targets wide-area MANET, is limited by the assumption of an online certification authority in the MANET context. Moreover, nodes could flood the network with packets destined to nonexistent nodes and possibly lead nodes unable to forward purchased packets to starvation. The practicality of the scheme is also limited by its assumptions, the high computational overhead (hop-byhop public key cryptography, for each transmitted packet), and the implementation of physically tamper-resistant modules.

The protection of the route discovery process has been regarded as an additional Quality-of-Service (QoS) issue [56], by choosing routes that satisfy certain quantifiable security criteria. In particular, nodes in a MANET subnet are classified into different trust and privilege levels. A node initiating a route discovery sets the sought security level for the route; the required minimal trust level for nodes participating in the query/reply propagation. Nodes at each trust level share symmetric encryption and decryption keys. Intermediate nodes of different levels cannot decrypt in-transit routing packets, or determine whether the required QoS parameter can be satisfied, and simply drop them. Although this scheme provides protection (e.g., integrity) of the routing protocol traffic, it does not eliminate false routing information provided by malicious nodes. Moreover, the proposed use of symmetric cryptography allows any node to corrupt the routing protocol operation within a level of trust, by mounting virtually any attack that would be possible without the presence of the scheme. Finally, the assumed supervising organization and the fixed assignment of trust levels does not pertain to the MANET paradigm. In essence, the proposed solution transcribes the problem of secure routing in a context, where nodes of a certain group are assumed to be trustworthy, without actually addressing the global secure routing problem.

An extension of the ad hoc on-demand distance vector (AODV) [18] routing protocol has been proposed [57] to protect the routing protocol messages. The secure AODV (S-AODV) scheme assumes that each node has certified public keys of all network nodes, so that intermediate nodes can validate all in-transit routing packets. The basic idea is that the originator of a control message appends an RSA signature [58] and the last element of a hash chain [59] (i.e., the result of nconsecutive hash calculations on a random number). As the message traverses the network, intermediate nodes cryptographically validate the signature and the hash value, generate the kth element of the hash chain, where k is the number of traversed hops, and place it in the packet. The route replies are provided either by the destination or intermediate nodes having an active route to the sought destination, with the latter mode of operation enabled by a different type of control packets.

The use of public key cryptography imposes a high processing overhead on the intermediate nodes and can be considered unrealistic for a wide range of network instances. Furthermore, it is possible for intermediate nodes to corrupt the route discovery by pretending that the destination is their immediate neighbor, advertising arbitrarily high sequence numbers and altering (either decreasing by one or arbitrarily increasing) the actual route length. Additional vulnerabilities stem from the fact that the IP portion of the S-AODV traffic can be trivially compromised, since it is not (and cannot be, due to the AODV operation) protected, unless additional hop-by-hop cryptography and accumulation of signatures is used. Finally, the assumption that certificates are bound with IP addresses is unrealistic; roaming nodes joining MANET sub-domains will be assigned IP addresses dynamically (e.g., DHCP [60]) or even randomly (e.g., zero configuration [61]).

A different approach is taken by the secure message transmission (SMT) [62] protocol, which, given a topology view of the network, determines a set of diverse paths connecting the source and the destination nodes. Then, it introduces limited transmission redundancy across the paths, by dispersing a message into N pieces, so that successful reception of any M out of N pieces allows the reconstruction of the original message at the destination. Each piece is equipped with a cryptographic header that provides integrity and replay protection, along with origin authentication, and is transmitted over one of the paths. Upon reception of a number of pieces, the destination generates an acknowledgment

² An alternative implementation, with each packet carrying a purse of fictitious currency from which nodes remove their reward, faces different challenges as well.

informing the source of which pieces, and thus routes, were intact. In order to enhance the robustness of the feedback mechanism, the small-sized acknowledgments are maximally dispersed (i.e., successful reception of at least one piece is sufficient) and are protected by the protocol header as well. If less than M pieces were received, the source retransmits the remaining pieces over the intact routes. If too few pieces were acknowledged or too many messages remain outstanding, the protocol adapts its operation, by determining a different path set, reencoding undelivered messages and reallocating pieces over the path set. Otherwise, it proceeds with subsequent message transmissions.

The protocol exploits MANET features such as the topological redundancy, interoperates widely with accepted techniques such as source routing, relies on a security association between the source and the destination, and makes use of highly efficient symmetric key cryptography. It does not impose processing overhead on intermediate nodes, while the end nodes make the routing decisions based on the feedback provided by the destination and the underlying topology discovery and route maintenance protocols. The fault tolerance of SMT is enhanced by the adaptation of parameters such as the number of paths and the dispersion factor (i.e., the ratio of required pieces to the total number of pieces). SMT can yield 100% successful message reception, even if 10-20% of the network nodes are malicious. Moreover, algorithms for the selection of path sets with different properties, based on different metrics and the network feedback, can be implemented by SMT. SMT provides a flexible, endto-end, secure traffic engineering scheme, tailored to the MANET characteristics.

It is noteworthy that SMT provides a limited protection against the use of compromised topological information, although its main focus is to safeguard the data forwarding operation. The use of multiple routes compensates for the use of partially incorrect routing information [52], rendering a compromised route equivalent to a route failure. Nevertheless, the disruption of the route discovery can still be the most effective way for adversaries to consistently compromise the communication of one or more pairs of nodes.

Another approach to secure the route discovery procedures, the secure routing protocol (SRP), has been proposed [63]. The scheme guarantees that a node initiating a route discovery will be able to identify and discard replies providing false topological information, or avoid receiving them. The novelty of the scheme, as compared with other MANET secure routing schemes, is that false route replies, as a result of malicious node behavior, are discarded partially by benign nodes while in transit toward the querying node, or deemed invalid on reception. The security goals are achieved with the existence of a security association between the pair of end nodes *only*, without the need for intermediate nodes to cryptographically validate control traffic.

The widely accepted technique in the MANET context of route discovery based on broadcasting query packets is the basis of the SRP protocol. More specifically, as query packets traverse the network, the relaying intermediate nodes append their identifier (e.g., *IP* address) in the query packet header. When one or more queries arrive at the sought destination, replies that contain the accumulated routes are returned to the querying node; the source then may use one or more of these routes to forward its data.

Reliance on this basic route query broadcasting mechanism allows SRP to be applied as an extension of a multitude of existing routing protocols. In particular, the dynamic source routing (DSR) [17] and the IERP [64] of the zone routing protocol (ZRP) [65] framework are two protocols that can be extended in a natural way to incorporate SRP. Furthermore, other protocols such as ABR [66], for example, could be combined with SRP with minimal modifications to achieve the security goals of the SRP protocol.

In SRP, only the end nodes have to be securely associated, and there is no need for cryptographic validation of control traffic at intermediate nodes; two factors that render the scheme efficient and scalable. SRP places the overhead on the end nodes, an appropriate choice for a highly decentralized environment, and contributes to the robustness and flexibility of the scheme. Moreover, SRP does not rely on the state stored in intermediate nodes, thus is immune to malicious acts not directed against the nodes that wish to communicate in a secure manner. Finally, SRP provides one or more route replies, whose correctness is verified by the route "geometry" itself.

6. SOME CONCLUDING THOUGHTS

The ad hoc networking technology has stimulated substantial research activity since the early 1990s. The rather interesting fact is that although the military has been experimenting and even using this technology since 1970s, the research community has been coping with the rather frustrating task of finding a "killer" nonmilitary application for ad hoc networks. A major challenge that has been perceived as a possible "show stopper" for technology transfer is the fact that commercial applications do not necessarily conform to the "collaborative" environment that the military communication environment does. In other words, why should a user forward someone else's transmission, depleting his/her own battery power and, thus, possibly restricting his/her use of the network in the future? This question may relate to the issue of billing - if billing is possible (and, in fact, desirable), then nodes that serve as "good citizens" could be rewarded. But billing is, by itself, a significant challenge in ad hoc networks.

Other challenges in deployment of ad hoc networks relate to the issues of manageability, security, and availability of communication through this type of technology.

Realizing that technology transfer has not been the motivating factor, the most recent research interest in ad hoc networks could be, most probably, attributed to the intellectual challenges that are part of this type of communication environment.

Nevertheless, we believe that there is substantial commercial potential of ad hoc networks. Future extensions of the cellular infrastructure could be carried out using this type of technology and may very well be the basis for the fourth-generation (4G) of wireless systems. Other possible applications include sensing systems (also referred to as *sensor networks*) or augmentation to the wireless LAN technology.

BIOGRAPHIES

Zygmunt J. Haas received his B.Sc. in EE in 1979 and M.Sc. in EE in 1985. In 1988, he earned his Ph.D. from Stanford University, Stanford, California, and subsequently joined AT&T Bell Laboratories in the Network Research Department. There he pursued research on wireless communications, mobility management, fast protocols, optical networks, and optical switching. From September 1994 to July 1995 Dr. Haas worked for the AT&T Wireless Center of Excellence, where he investigated various aspects of wireless and mobile networking, concentrating on TCP/IP networks. As of August 1995, he joined the faculty of the School of Electrical and Computer Engineering at Cornell University, Ithaca, New York.

Dr. Haas is an author of numerous technical papers and holds 15 patents in the fields of high-speed networking, wireless networks, and optical switching. He has organized several workshops, delivered tutorials at major IEEE and ACM conferences, and serves as editor of several journals and magazines, including the IEEE Transactions on Networking. He has been a guest editor of three IEEE JSAC issues ("Gigabit Networks," "Mobile Computing Networks," and "Ad-Hoc Networks"). Dr. Haas is a senior member of IEEE, a voting member of ACM, and the chair of the IEEE Technical Committee on Personal Communications. His interests include: mobile and wireless communication and networks, personal communication service, and high-speed communication and protocols. His e-mail is: haas@ece.cornell.edu and his URL is: http://wnl.ece.cornell.edu.

Jing Deng received his B.E. in 1994 and M.E. in 1997, both in EE, from Tsinghua University in Beijing, P. R. China. He earned his Ph.D. degree from Cornell University, Ithaca, New York, in 2002. Dr. Deng's interests include protocol design, evaluation, and optimization for mobile ad-hoc networks. His e-mail address is: jing@ece.cornell.edu.

Dr. Ben Liang received his B.Sc. and M.Sc. degrees, both in electrical engineering, from Polytechnic University in 1997. In 2001, he received his Ph.D. degree in electrical engineering from Cornell University, Ithaca, New York. He is currently a postdoctoral research associate and visiting lecturer in the School of Electrical and Computer Engineering at Cornell University. He has pursued research in wireless networking, mobility management, distributed fault-tolerance, and image processing. His current research interests are in developing theories, algorithms, and protocols for wireless and mobile communication systems. Dr. Liang is a member of IEEE and Tau Beta Pi.

Panagiotis Papadimitratos is a Ph.D. candidate in the Electrical and Computer Engineering Department at

Cornell University, Ithaca, New York. He is currently a graduate research assistant, member of the Wireless Network Laboratory, working under the supervision of Professor Z. J. Haas. His research examines issues in the general areas of mobile and wireless networking. His work focuses on the design and evaluation of secure and fault tolerant routing protocols to support mobile ad-hoc networking. Mr. Papadimitratos joined the Ph.D. program at Cornell University in 1998, having received his Diploma degree from the Computer Engineering and Informatics Department at the University of Patras, Patras, Greece. His Diploma thesis examined issues on adaptive and blind equalization for cellular mobile networks.

S. Sajama received her bachelor of technology from the Indian Institute of Technology, Bombay, India, in 1998. Subsequently, she has joined the School of Electrical and Computer Engineering at Cornell University, Ithaca, New York, obtaining her masters degree in August 2001. Her thesis concentrated on multicast routing protocols with application to ad-hoc networks. Her e-mail address is: sajama@ece.cornell.edu

BIBLIOGRAPHY

- 1. C. E. Perkins, ed., Ad Hoc Networking, Addison-Wesley Longman, 2001.
- F. A. Tobagi and L. Kleinrock, Packet switching in radio channels: Part II—The hidden terminal problem in carrier multiple-access and the busy-tone solution, *IEEE Trans. Commun.* COM-23(12): 1417-1433 (Dec. 1975).
- J. Postel, Internet Control Message Protocol, RFC 792, IETF, Sept. 1981.
- P. Karn, MACA—a new channel access method for packet radio, ARRL/CRRL Amateur Radio 9th Computer Networking Conf., 1990, pp. 134–140.
- 4.5 Wu C. and Li V. O. K., Receiver-initiated busy-tone multiple access in packet radio networks, *Proc. ACM SIGCOMM*'87, 1987, pp. 336-342.
- V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, MACAW: A media access protocol for wireless LAN's, *Proc.* ACM SIGCOMM'94, 1994, pp. 212–225.
- C. L. Fullmer and J. J. Garcia-Luna-Aceves, Floor acquisition multiple access (FAMA) for packet-radio networks, *Proc. ACM* SIGCOMM'95, 1995, pp. 262–273.
- C. L. Fullmer and J. J. Garcia-Luna-Aceves, Solutions to hidden terminal problems in wireless networks, *Proc. ACM* SIGCOMM'97, 1997, pp. 39–49.
- 8. Z. J. Haas and J. Deng, Dual busy tone multiple access (DBTMA): A multiple access control scheme for ad hoc networks, *IEEE Trans. Commun.* (in press).
- 9. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice-Hall, 1992.
- C. Cheng, R. Reley, S. P. R. Kumar, and J. J. Garcia-Luna-Aceves, A loop-free extended Bellman-Ford routing protocol without bouncing effect, ACM Comput. Commun. Rev. 19(4): 224-236 (1989).
- J. J. Garcia-Luna-Aceves, Loop-free routing using diffusing computations, *IEEE/ACM Trans. Network.* 1(1): 130-141 (Feb. 1993).

- C. E. Perkins and P. Bhagwat, Highly dynamic destinationsequenced distance-vector routing (DSDV) for mobile computers, ACM SIGCOMM Comput. Commun. Rev. 24(4): 234–244 (Oct. 1994).
- P. Jacquet, P. Muhlethaler, and A. Qayyum, *Optimized Link State Routing Protocol*, IETF MANET, Internet Draft, Nov. 1998.
- S. Murthy and J. J. Garcia-Luna-Aceves, A routing protocol for packet radio networks, Proc. ACM Mobile Computing and Networking Conf., MOBICOM'95, Nov. 14–15, 1995.
- S. Murthy and J. J. Garcia-Luna-Aceves, An efficient routing protocol for wireless networks, ACM Mobile Networks Appl. J. 1(2): 187-197 (Oct. 1996).
- V. D. Park and M. S. Corson, A highly adaptive distributed routing algorithm for mobile wireless networks, *IEEE INFOCOM* '97, Kobe, Japan, 1997.
- D. B. Johnson and D. A. Maltz, Dynamic source routing in ad hoc wireless networks, in T. Imielinski and H. Korth, eds., *Mobile Computing*, Kluwer Academic Publishers, 1996, pp. 153-181.
- C. E. Perkins and E. M. Royer, Ad-hoc on-demand distance vector routing, Proc. 2nd IEEE Workshop on Mobile Computing Systems and Applications, Feb. 1999, pp. 90-100.
- M. R. Pearlman and Z. J. Haas, Determining the optimal configuration for the zone routing protocol, *IEEE J. Select. Areas Commun.* 17(8): 1395-1414 (Aug. 1999).
- V. Park and M. S. Corson, IETF MANET Internet Draft, draft-ietf-MANET-tora-spec-03.txt, Nov. 2000.
- E. Gafni and D. Bertsekas, Distributed algorithms for generating loop-free routes in networks with frequently changing topology, *IEEE Trans. Commun.* 29(1): 11–15 (Jan. 1981).
- 22. M. S. Corson and A. Ephremides, A distributed routing algorithm for mobile wireless networks, ACM/Baltzer Wireless Networks 1(1): 61–81 (Feb. 1995).
- P. Jacquet et al., IETF MANET Internet Draft, draft-ietf-MANET-olsr-02.txt, July 2000.
- A. Iwata et al., Scalable routing strategies for ad hoc wireless networks, *IEEE J. Select. Areas Commun.* 17(8): 1369–1379 (Aug. 1999).
- T.-W. Chen and M. Gerla, Global state routing: A new routing scheme for ad-hoc wireless networks, *IEEE ICC* 171–175 (June 1998).
- R. Sivakumar, P. Sinha, and V. Bharghavan, CEDAR: A coreextraction distributed ad hoc routing algorithm, *IEEE J. Select. Areas Commun.* 17(8): 1454–1465 (Aug. 1999).
- M. Joa-Ng and I.-T. Lu, A peer-to-peer zone-based two-level link state routing for mobile ad hoc networks, *IEEE J. Select. Areas Commun.* 17(8): 1415-1425 (Aug. 1999).
- P. F.Tsuchiya, The landmark hierarchy: A new hierarchy for routing in very large networks, *Comput. Commun. Rev.* 18(4): 35-42 (Aug. 1988).
- 29. G. Pei, M. Gerla, and X. Hong, LANMAR: Landmark routing for large scale wireless ad hoc networks with group mobility, *1st Annual Workshop on Mobile and Ad Hoc Networking and Computing (MobiHOC)*, Aug. 2000.
- Y.-B. Ko and N. H. Vaidya, Location-aided routing (LAR) in mobile ad hoc networks, ACM/IEEE MobiCom, Dallas, TX, 1998.

- S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, A distance routing effect algorithm for mobility (DREAM), ACMIEEE MobiCom, Dallas, TX, 1998.
- S. Paul, Multicasting on the Internet and its Applications, Kluwer, 1998.
- C.-H. Chow, On multicast path finding algorithms, Proc. IEEE INFOCOM'91, 1991, pp. 1274-1283.
- B. M. Waxman, Performance evaluation of multipoint routing algorithms, Proc. IEEE INFOCOM'93, 1993, pp. 980–986.
- B. Kadaba and J. M. Jaffe, Routing to multiple destinations in computer networks, *IEEE Trans. Commun.* COM-31(3): 343-351 (March 1983).
- V. P. Kompella, J. C. Pasquale, and G. C. Polyzos, Multicast routing for multimedia communication, *IEEE/ACM Trans. Network.* 1(3): 286-292 (June 1993).
- S. E. Deering et al., An architecture for wide-area multicast routing, *IEEE/ACM Trans. Network.* 4(2): 153-162 (April 1996).
- A. Ballardie, Core Based Trees (CBT Version 2) Multicast Routing—Protocol Specification, RFC-2189, Sept. 1997.
- C. Chiang, M. Gerla, and L. Zhang, Shared tree wireless network multicast, *IEEE International Conf. Computer Communications and Networks (ICCCN'97)*, Sept. 1997.
- S.-J. Lee, M. Gerla, and C.-C. Chiang, On-demand multicast routing protocol, *IEEE WCNC'99*, New Orleans, LA, Sept. 1999, pp. 1298–1304.
- C. W. Wu and Y. C. Tay, AMRIS: A multicast protocol for ad hoc wireless networks, *IEEE MILCOM'99*, Atlantic City, NJ, Nov. 1999.
- 42. J. J. Garcia-Luna-Aceves and E. L. Madruga, The Coreassisted mesh protocol, *IEEE J. Select. Areas Commun.* 17(8): (Aug. 1998).
- E. Royer and C. E. Perkins, Multicast operation of ad-hoc, on-demand distance vector routing protocol, ACM/IEEE MobiCom'99, Aug. 1999.
- E. Bommaiah, A. McAuley, R. Talpade, and M. Liu, AMRoute: Ad hoc multicast routing protocol, Internet-Draft, IETF, Aug. 1998.
- 45. S. Lee and C. Kim, Neighbor supporting ad hoc multicast routing protocol, 2000 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing, 2000, pp. 37–44.
- 46. L. Briesemeister and G. Hommel, Role-based multicast in highly mobile but sparsely connected ad hoc networks, 2000 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing, 2000, pp. 45–50.
- 47. H. Zhou and S. Singh, Content based multicast in ad hoc networks, 2000 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing, 2000, pp. 51-60.
- 48. G. D. Kondylis, S. V. Krishnamurthy, S. K. Dao, and G. J. Pottie, Multicasting sustained CBR and VBR traffic in wireless ad hoc networks, *Int. Conf. Communications*, 2000, pp. 543–549.
- C. Sankaran and A. Ephremides, Multicasting with multiuser detection in ad-hoc wireless networks, Conf. Information Sciences and Systems (CISS), 1998, pp. 47-54.
- 50. P. Sinha, R. Sivakumar, and V. Bharghavan, MCEDAR: Multicast core-extraction distributed ad hoc routing, Wireless Communications and Networking Conf., 1999, pp. 1313-1318.

2898 WIRELESS AD HOC NETWORKS

- 51. J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, Algorithms for energy-efficient multicasting in ad hoc wireless networks, *IEEE Military Communications Conf. (MILCOM)*, 1999, pp. 1414–1418.
- L. Zhou and Z. J. Haas, Securing ad hoc networks, *IEEE Network Mag.* 13(6): (Nov./Dec. 1999).
- 53. F. Stajano and R. Anderson, The resurrecting duckling: Security issues for ad hoc wireless networks, *Security Protocols*, 7th Int. Workshop, LNCS, Springer-Verlag, 1999.
- 54. S. Marti, T. J. Giuli, K. Lai, and M. Baker, Mitigating routing misbehavior in mobile ad hoc networks, 6th MOBICOM Conf., Boston, Aug. 2000.
- 55. L. Buttyan and J. P. Hubaux, Enforcing service availability in mobile ad hoc WANs, *1st MobiHoc conf.*, Boston, Aug. 2000.
- 56. S. Yi, P. Naldurg, and R. Kravets, Security-Aware Ad-Hoc Routing for Wireless Networks, UIUCDCS-R-2001-2241 Technical Report, Aug. 2001.
- 57. M. Guerrero, Secure AODV, Internet draft sent to *manet@itd.nrl.navy.mil* mailing list, Aug. 2001.
- R. Rivest, A. Shamir, and L. Adleman, A method for obtaining Digital Signatures and public key cryptosystems, *Commun.* ACM 21(2): 120-126 (Feb. 1978).
- L. Lamport, Password authentication with insecure communication, Commun. ACM 24(11): 770-772 (Nov. 1981).
- 60. R. Droms, Dynamic Host Configuration Protocol, RFC 2131, IETF, March 1997.
- 61. M. Hattig, ed., Zero-conf IP host requirements, *draft-ietf-zeroconf-reqts-09.txt*, IETF MANET Working Group, Aug. 31, 2001.
- 62. P. Papadimitratos and Z. J. Haas, Secure message transmission in mobile ad hoc networks (in preparation).
- 63. P. Papadimitratos and Z. J. Haas, Secure routing for mobile ad hoc networks, SCS Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS 2002), San Antonio, TX, Jan. 27–31 2002.
- 64. Z. J. Haas, M. Perlman, and P. Samar, The interzone routing protocol (IERP) for ad hoc networks, *draft-ietf-manet-zoneierp-01.txt*, MANET Working Group, IETF, June 1, 2001.
- 65. Z. J. Haas and M. Perlman, The performance of query control schemes of the zone routing protocol, *IEEE/ACM Trans. Network.* 9(4): 427–438 (Aug. 2001).
- C. K. Toh, Associativity-based routing for ad-hoc mobile networks, Wireless Pers. Commun. 4(2): 1–36 (March 1997).

FURTHER READING

- Asokan N. and P. Ginzboorg, Key agreement in ad hoc networks, Comput. Commun. 23(17): 1627–1637 (Nov. 2000).
- Bellovin S. M. and M. Merritt, Encrypted key exchange: Password-based protocols secure against dictionary attacks, *IEEE* SympSecurity and Privacy (May 1992).
- Bellur B. and R. G. Ogier, A reliable, efficient topology broadcast protocol for dynamic networks, *IEEE INFOCOM*, New York, March 1999.
- Bianchi G., Performance analysis of the IEEE 802.11 distributed coordination function, *IEEE JSAC* 18(3): 535–547 (March 2000).
- Binkley J. and W. Trost, Authenticated ad hoc routing at the link layer for mobile systems, *ACM/Baltzer Wireless Networks* 139–145 (March 2001).

- Blaze M., J. Feigenbaum, J. Ioannidis, and A. D. Keromytis, *The KeyNote Trust-Management System*, RFC 2074, IETF, Sept. 1999.
- Broch J., D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, A performance comparison of multi-hop wireless ad hoc network routing protocols, ACM/IEEE MobiCom, 1998, pp. 85–97.
- Chiang C.-C., H.-K. Wu, W. Liu, and M. Gerla, Routing in clustered multihop, mobile wireless networks with fading channel, *IEEE Singapore Int. Conf. on Networks*, 1997.
- Das S. R., C. E. Perkins, and E. M. Royer, Performance comparison of two on-demand routing protocols for ad hoc networks, *IEEE INFOCOM* 1: 3–12 (March 2000).
- Dube R., C. D. Rais, K.-Y. Wang, and S. K. Tripathi, Signal stability-based adaptive routing (SSA) for ad hoc mobile networks, *IEEE Pers. Commun.* 4(1): 36-45 (Feb. 1997).
- Ephremeides A., J. E. Wieselthier, and D. J. Baker, A design concept for reliable mobile radio networks with frequency hopping signaling, *Proc. IEEE* **75**(1): (Jan. 1987).
- Feeney L. M., B. Ahlgren, and A. Westerlund, Spontaneous networking: An application-oriented approach to ad hoc networking, *IEEE Commun. Mag.* **39**(6): 176–181 (June 2001).
- FedInfProcStandards, Secure Hash Standard, Publication 180, NIST, May 1993.
- Gerla M. and T. C. Tsai, Multiuser, mobile multimedia radio network, ACM/Balzer J. Wireless Networks 1(3): 255–265 (1995).
- Gerla M., K. Tang, and R. Bagrodia, TCP performance in wireless multi-hop networks, *Mobile Comput. Syst. Appl.* 41–50, 25–26 (Feb. 1999).
- Garcia-Luna-Aceves J. J. and M. Spohn, Source-tree routing in wireless networks, *IEEE ICNP 99: 7th Int. Conf. Network Protocols*, Toronto, Canada, Oct. 1999.

http://www.scienceforum.com/glomo/.

- Hauser R., T. Przygenda, and G. Tsudik, Lowering security overhead in link state routing, *Comput. Networks* 31: 885–894 (1999).
- IEEE STD 802.11, Wireless LAN Media Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE, 1999.
- Jiang M., J. Li, and Y. C. Tay, IETF MANET Internet Draft, draft-ietf-MANET-cbrp-spec-01.txt, Aug. 1999.
- Johansson P. et al., Scenario-based performance analysis of routing protocols for mobile ad-hoc networks, ACM/IEEE MobiCom, Aug. 1999, pp. 195-206.
- Jubin J. and J. D. Tornow, The DARPA packet radio network protocols, Proc. IEEE 75: 21–32 (Jan. 1987).
- Kershenbaum A. and W. Chou, A unified algorithm for designing multidrop teleprocessing networks, *IEEE Trans. Commun.* COM-22: 1762–1772 (Nov. 1974).
- Kondylis G. D., S. V. Krishnamurthy, S. K. Dao, and G. J. Pottie, Multicasting sustained CBR and VBR traffic in wireless ad hoc networks, *Int. Conf. Communications*, 2000, pp. 543–549.
- Nikaein N., H. Labiod, and C. Bonnet, DDR-distributed dynamic routing algorithm for mobile ad hoc networks, 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing (MobiHOC), Aug. 2000.
- Perkins C., IP Mobility Support, RFC 2002, IETF, Oct. 1996.
- Ramanathan R. and M. Steenstrup, Hierarchically-organized, multihop mobile wireless networks for quality-of-service support, ACM/Baltzer Mobile Networks Appl. 3(1): 101–119 (1998).
- Royer E. M. and C.-K. Toh, A review of current routing protocols for ad hoc mobile wireless networks, *IEEE Pers. Commun.* (April 1999).

- Santivanez C., Asymptotic Behavior of Mobile Ad Hoc Routing Protocols with Respect to Traffic, Mobility and Size, Technical Report TR-CDSP0052, Dept. Electrical and Computer Engineering, Northeastern Univ., Boston, 2000.
- Shacham N. and J. Westcott, Future directions in packet radio architectures and protocols, *Proc. IEEE* **75**: 83-99 (Jan. 1987).
- Sharony J., An architecture for mobile radio networks with dynamically changing topology using virtual subnets, ACM Mobile Networks Appl. J. 1(1): 75-86 (1996).
- Stajano F., The resurrecting duckling—what next? Security Protocols, 8th International Workshop, LNCS, Springer-Verlag, 2000.
- Xu S. and T. Saadawi, Does the IEEE 802.11 MAC Protocol Work Well in Multihop Wireless Ad Hoc Networks? *IEEE Communi. Magazine* 130–137 (June 2001).

WIRELESS APPLICATION PROTOCOL (WAP)

ALESSANDRO ANDREADIS GIOVANNI GIAMBENE University of Siena Siena, Italy

1. INTRODUCTION

It is expected that within a few years the number of mobile devices accessing the Internet will exceed the number of *personal computers* (PCs). The use of mobile terminals is becoming attractive, since they allow the access on the move and require a shorter setup time than the initial power-on of a PC. However, Internet services have not been developed for mobile devices, as they are not suitable for small displays and they are not personalized or locationdependent. A solution to these needs is represented by the *Wireless Application Protocol* (WAP), which is an open, global specification that empowers mobile users with wireless devices for easy Internet access [1,2].

WAP is the result of the WAP Forum's efforts to promote industrywide specifications for developing applications and services that operate over wireless communication networks [3]. The WAP Forum was formed after a U.S. network operator, Omnipoint, issued a tender for the supply of mobile information services in early 1997. It received several responses from different suppliers using proprietary techniques for delivering the information such as Smart Messaging from Nokia and Handheld Device Markup Language (HDML) from Phone.com. Omnipoint informed the tender responders that it would not accept a proprietary approach and recommended that various vendors get together to explore defining a common standard. After all, there was not a great deal of difference between the different approaches, which could be combined and extended to form a powerful standard. These events triggered the development of WAP, with Ericsson and Motorola joining Nokia and Phone.com as the founder members of the WAP Forum. At present, the WAP Forum encompasses more than 500 members.

WAP is the de facto standard for porting Internet services to wireless devices, such as mobile phones and Personal Digital Assistants (PDA). WAP contains a lightweight protocol stack (based on the Internet one) well suited for the wireless scenario and an application environment that allows delivering information services to mobile phones. WAP can be built on many mobile phone operating systems, including: PalmOS, EPOC, Windows CE and JavaOS [3]. WAP also contains a microbrowser according to which the information received is interpreted in the handset and presented to the user. WAP is designed to work with most wireless networks such as CPDC, CDMA, GSM, PDC, PHS, TETRA, and DECT [3]. WAP devices, despite the current rather limited user interface, provide a valuable means to access corporate and public services. Examples of applications are

- Use of a corporate application in less time than it takes to boot the laptop
- Availability of the device on the move
- Access to services from several countries
- Mobile browsing and mobile access to email

2. WAP SYSTEM BUILDING BLOCKS

The WAP network architecture envisages both WAP servers, hosting pages designed in a suitable markup language and WAP gateways between the wireless network and the wireline Internet (see Fig. 1) [4]. WAP 1.0 was based on *Wireless Markup Language* (WML), a subset of the *eXtensible Markup Language* (XML). The basic markup language in WAP 2.0, namely, WML2, is based on the *eXtensible HyperText Markup Language* (XHTML), as defined by the *World Wide Web Consortium* (W3C) [5]. By using the XHTML modularization approach, the WML2 language is very extensible, permitting additional language elements to be added as needed.

The WAP protocol architecture is based on a client/server model, as sketched in Fig. 2, where we can see the mobile client, that is the mobile user, the WAP proxy and the Web server. The client Web browser makes a request for a Webpage. This request is sent to the WAP proxy that acts as a gateway for the Internet [6]. Through a protocol conversion, a *HyperText Transfer Protocol* (HTTP) request is thus sent to the appropriate Web



Figure 1. WAP system architecture.



Figure 3. WML compiling and encoding.

server. The response is a bytestream of ASCII text, which is a *HyperText Markup Language* (HTML) Webpage. The use of *Common Gateway Interface* (CGI) programs or Java servlets allows for the dynamic creation of HTML pages using content stored in a database. The Webpage is sent to the WAP proxy that first translates it from HTML to the WML language (see Fig. 3, step 2). Finally, a page conversion is performed in a compact binary representation that is suitable for wireless networks (see Fig. 3, step 3) [7]. In the case that there is a WAP server (i.e., hosting pages in the WML format) in the mobile network, the mobile client directly receives WML pages from the server without involvement of the WAP gateway.

The infrastructure required to deliver WAP services to the mobile terminal is similar to that of the existing WWW model. The Web server is the same product; the one most commonly used is *Apache*. The Web server needs to be configured to serve the pages written in WML as well as HTML.

Although reusing of existing Internet content by means of on-the-fly adaptations and translations is an explicit goal, test realizations of the WAP gateway and proxy servers show that the creation of new content that is explicitly designed for presentation using WML is a more effective option.

Since a mobile user cannot use a QWERTY keyboard or a mouse, WML documents are structured into a set of well-defined units of user interactions called *cards*. Each card may contain instructions for gathering user input, information to be presented to the user, and similar (see Fig. 4). A single collection of cards is called a *deck*, which is the unit of content transmission, identified by



Figure 4. Internal organization of a WML page (=deck) in cards, with different tags.

a *Uniform Resource Locator* (URL) [8]. After browsing a deck, the WAP-enabled phone displays the first card; then, the user decides whether to proceed to the next card of the same deck. WML content is scalable from a two-line text display on a basic device to a full graphic screen on the latest smart phones and communicators. WML supports:

- Text (bold, italics, underlined, line breaks, tables)
- Black-and-white images (wireless bitmap format, WBMP)
- User input
- Variables
- Navigation and history stack
- Scripting (WMLScript), a lightweight script language, similar to JavaScript

In particular, WML includes support for managing user agent state by means of variables and for tracking the history of the interaction. Moreover, WMLScripts are sent separately from decks and are used to enhance the client *man-machine interface* (MMI) with sophisticated device and peripheral interactions.

The Wireless Telephony Application (WTA) of WAP contains a client-side WTA programming library and a WTA server (see Figs. 1 and 2); together they allow the WAP session to control the voice channel. WTA and its interface, WTA-Interface (WTAI), provide the access and the programming interface to telephony services. The WTA server generates WTA events interpreted by the WAP gateway that sends the resulting WML to the WAP mobile phone. The WTA server then initiates and controls any voice connections that are required.

3. WAP PROTOCOL STACK

WAP protocol and its functions are layered similarly to the OSI Reference Model [9]. In particular, the WAP protocol stack is analogous to the Internet one (see Fig. 5). Each layer is accessible by the layers above, as well as by other services and applications. The WAP layered architecture enables other services and applications to utilize the features of the WAP stack through a set of well-defined interfaces. Figure 6 compares Internet and WAP protocol stacks. A brief survey of the protocols at the different WAP layers is provided below.

3.1. Wireless Application Environment (WAE)

WAE specifies an application framework for wireless devices such as mobile phones, pagers, and PDAs. WAE specifies the markup languages and acts as a container for applications such as a microbrowser. In particular, WAE encompasses the following parts:

- WML Microbrowser
- WMLScript Virtual Machine
- WMLScript Standard Library
- Wireless Telephony Application Interface (i.e., telephony services and programming interfaces)
- WAP Content Types

The two most important formats defined in WAE are the WML and WMLScript byte-code formats. A WML encoder at the WAP gateway, or "tokenizer," converts a WML deck into its binary format (see Fig. 3, step 3) [7] and a WMLScript compiler takes a script into byte-code. This process allows a significant compression of the data to be transmitted on the air interface, thus making more efficient the transmission of WML and WMLScript data.



Figure 6. Internet and WAP 1.0 protocol

stacks

Figure 5. WAP 1.0 protocol stack.

3.2. Wireless Session Protocol (WSP)

WSP provides the application layer of WAP (i.e., WAE) with a consistent interface for two-session services. The first is a connection-oriented service above the *Wireless Transaction Protocol* (WTP). The second is a connectionless service operating above a secure or nonsecure datagram service [*Wireless Datagram Protocol* (WDP)]. WSP is the equivalent of the HTTP protocol in both the Internet and WAP 2.0 release that supports the TCP/IP levels in the protocol stack.

3.3. Wireless Transaction Protocol (WTP)

WTP runs on top of a datagram service [such as the *User Datagram Protocol* (UDP)] and provides a lightweight transaction-oriented protocol that is suitable for implementation in mobile terminals. WTP offers three classes of transaction services: unreliable one-way request, reliable one-way request and reliable two-way request respond.

3.4. Wireless Transport Layer Security (WTLS)

WTLS is a security protocol based on the industrystandard *Transport Layer Security* (TLS) protocol. WTLS is intended for use with the WAP transport protocols and has been optimized for wireless communication networks. It includes data integrity checks, privacy on the WAP gateway-to-client leg, and authentication.

3.5. Wireless Datagram Protocol (WDP)

WDP is transport-layer protocol in WAP [10]. WDP supports connectionless reliable transport and bearer independence. WDP offers consistent services to the upper-layer protocols of WAP and operates above the data-capable bearer services supported by various air interface types. Since the WDP protocols provide a common interface to upper-layer protocols, the security, session, and application layers are able to operate independently of the underlying wireless network. At the mobile terminal, the WDP protocol consists of the common WDP elements plus an adaptation layer that is specific for the adopted air interface bearer. The WDP specification lists the bearers that are supported and the techniques used to allow WAP protocols to operate over each bearer [3]. The WDP protocol is based on UDP. UDP provides port-based addressing, and IP provides *Segmentation And Reassembly* (SAR) in a connectionless datagram service. When the IP protocol is available over the bearer service, the WDP datagram service offered for that bearer will be UDP.

3.6. Bearers on the Air Interface

Let us refer to the *Global System for Mobile communications* (GSM) network, where the following bearer services can be adopted to support WAP traffic [11]:

- Unstructured Supplementary Services Data (USSD)
- circuit-switched Traffic CHannel (TCH)
- Short Message Service (SMS)
- General Packet Radio Service (GPRS), plain data traffic
- Multimedia Messaging Service (MMS) over GPRS

Let us compare these different options to support WAP traffic. TCH has the disadvantage of a 30–40s connection

delay between the WAP client and the gateway, thus making it less suitable for mobile subscribers. SMS and USSD are inexpensive bearers for WAP data with respect to TCH, leaving the mobile device free for voice calls. SMS and USSD are transported by the same air interface channels. SMS is a store-and-forward service that relies on a Short Message Service Center (SMSC), whereas USSD is a connection-oriented (no store-andforward) service, where the Home Location Register (HLR) of the GSM network receives and routes messages from/to the users. The SMS bearer is well suited for WAP push applications (available from WAP release 1.2), where the user is automatically notified each time an event occurs. USSD is particularly useful for supporting transactions over WAP. Finally, GPRS radio transmissions allow a high capacity [<170 kbps (kilobits per second) using all the slots of a GSM carrier with the lightweight coding scheme] that is shared among mobile phones according to a packet-switching scheme. Hence, GPRS can provide an efficient scheme for WAP contents delivery.

The WAP protocol layers at the client, at the gateway, and at the Web server are detailed in Fig. 7. Figure 8 gives further details on different WAP protocol stack possibilities on the client side. In particular, the leftmost stack represents a typical example of a WAP application, namely, a WAE user agent running over the complete



Figure 7. WAP 1.0 protocol architecture at different interfaces.



Figure 8. Different possibilities for the WAP protocol stack on the client side.

2904 WIRELESS APPLICATION PROTOCOL (WAP)

portfolio of WAP technology. The middle stack is intended for applications and services that require transaction services with or without security. The rightmost stack is intended for applications and services that only require datagram transport with or without security.

4. COMPARISON BETWEEN WAP PROTOCOL RELEASES

The differences between the different releases of the WAP protocol are detailed below:

- WAP 1.0: first version of software for mobile clients, first adoption of WML, WBMP image format
- WAP 1.1: WTAI—"clickable" phone numbers, support of tables, boldface types, encrypted communication
- WAP 1.2: support of push applications, telephone identification, certificate handling
- WAP 2.0: latest WAP release, WML replaced by XHTML, colour screens, banners, MP3 and MP4 audio files, Internet radio, Bluetooth, remote control, integration with *Mobile Positioning System* (MPS) for locating the users (location-aware services)

5. TOOLS AND APPLICATIONS

The WAP programming model is similar to the WWW programming one. This fact provides several benefits to the application developer community, including a proven architecture and the ability to leverage existing tools (e.g., Web servers, XML tools). Optimizations and extensions have been made in order to match the characteristics of the wireless environment. Different WAP browsers can be found in Ref. 12; they are useful tools for developing WAPbased services for mobile users. WAP allows customers to easily reply to incoming information on the phone by adopting new menus to access mobile services.

Existing mobile operators have added WAP support to their offering by either developing their own WAP interface or, more often, partnering with one of the WAP gateway suppliers. WAP has also given new opportunities to allow the mobile distribution of existing information contents. For example, CNN and Nokia teamed up to offer CNN Mobile. Moreover, Reuters and Ericsson teamed up to provide Reuters Wireless Services.

New mobile applications that can be made available through a WAP interface include:

- Location-aware services
- · Web browsing
- Remote local-area network access
- Corporate email
- Document sharing/collaborative working
- Customer service
- Remote monitoring such as meter reading
- Job dispatch
- Remote point of sale
- File transfer

- Home automation
- Home banking and trading on line

Another group of important applications are based on the WAP push service that allows contents to be sent or "pushed" to devices by server-based applications via a push proxy. Push functionality is especially relevant for realtime applications that send notifications to their users, such as messaging, stock prices, and traffic update alerts. Without push functionality, these types of applications would require the devices to poll application servers for new information or status. In cellular networks such polling activities would cause an inefficient and wasteful use of the resources. WAP push functionality provides control over the lifetime of pushed messages, store-andforward capabilities at the push proxy, and control over the bearer choice for delivery.

Interesting WAP applications are made possible by the creation of dynamic WAP pages by means of the following different options:

- Microsoft ASP
- Java and servlets or *Java Server Pages* (JSPs) for generating WAP decks
- XSL Transformation (XSLT) for generating WAP pages adapted for displays of different characteristics and sizes

Alternative approaches to the use of WAP for mobile applications could be as follows:

- Subscriber Identity Module (SIM) Toolkit—the use of SIMs or smart cards in wireless devices is already widespread.
- Windows CE a multitasking, multithreaded operating system from Microsoft designed for including or embedding mobile and other space-constrained devices.
- JavaPhone—Sun Microsystems is developing PersonalJava and a JavaPhone Application Programming Interface (API), which is embedded in a Java virtual machine on the handset. Thus, cellular phones can download extra features and functions over the Internet.

SIM Toolkit and Windows CE are present days technologies as well as WAP. SIM Toolkint implies the definition of a set of services "embedded" on the SIM that allow users to contact several service providers through the mobile phone network. The Windows CE solution is based on an operating system developed for mobile devices and that may support different applications. Finally, JavaPhone will be the most sophisticated option for the development of device-independent applications.

Within the European Telecommunications Standards Institute (ETSI) and 3rd Generation Partnership Project (3GPP), standardization activities are in progress for the realization of mobile services. Accordingly, a new standard, called Mobile station application Execution Environment (MExE), has been defined [13]. In order to
ensure the portability of a variety of applications, across a broad spectrum of multivendor mobile terminals, a dynamic and open architecture has been standardized for both the Mobile Station (MS) and the SIM, that is a common set of APIs and development tools. MExE is based on the idea to specify a terminal-independent execution environment on the client device (i.e., MS and SIM) for nonstandardized applications and to implement a mechanism that allows the negotiation of supported capabilities (taking into account available bandwidth, display size, processor speed, memory, MMI). The key concept of the MExE service environment to make applications mobile-aware (i.e., aware of MS capabilities, network bearer characteristics, and user preferences) is the introduction of MExE classmarks that have been standardized as follows:

- *MExE classmark 1*—service based on WAP; requires limited input and output facilities (e.g., as simple as a 3-line × 15-character display and a numeric keypad) on the client side and is designed to provide quick and cheap information access even over narrow and slow data connections.
- *MExE classmark* 2—service based on PersonalJava; provides and utilizes a run-time system requiring more processing, storage, display, and network resources, but supports more powerful applications and more flexible MMIs. MExE classmark 2 also includes support for MExE classmark 1 applications (via the WML browser).

6. CONCLUSIONS

With the advent of the information society there is a growing need for network operators to support the mobile access to the Internet and its most popular applications such as Web browsing, email, file transfer, and remote login. The WAP protocol proposed by the WAP Forum is a first solution for allowing mobile access to the Internet. Despite its limitations, due to both the use of inadequate radio bearers (i.e., circuit-switched traffic channels) and the inefficient translation from HTML to WML (WAP 1 releases), WAP permits the mobile provision of services and contents. WAP can make available to users many information services that will be adequately supported by future-generation packet-switched bearers on the air interface.

BIOGRAPHIES

Alessandro Andreadis is assistant professor at the Department of Information Engineering of Siena University, Italy, since 1998. In 1993, he received the graduate degree in electronic engineering at the University of Florence, Italy. In the same year he won a research grant at the public administration of Regione Toscana, for a two-year research program on broadband networks based on SMDS and DQDB protocols. His work was funded for two further years, toward the development and diffusion of telematic services, via MAN networks, to small and medium enterprises of the territory. He held the courses of "Systems and Technologies for Communications" at the Department of Communication Science (University of Siena) and of "Telecommunication Networks" at the Faculty of Engineering (University of Siena). Here, he is presently teaching the course on transmission and processing of information in multimedia systems. Since 1995, he has been working at various international projects funded by the European Commission, in the Advanced Communication Technologies and Services (ACTS) Information Society Technologies (e IST) programs. His research interests focus on adaptive multimedia applications for mobile environments, traffic modeling, WAP services, TCP/IP on wireless and mobile networks.

Giovanni Giambene received the Dr. Ing. degree in electronics and a Ph.D. degree in telecommunications and informatics from the University of Florence, Italy, in 1993 and in 1997, respectively. From 1994 to 1996 he was technical external secretary of the European Community project COST 227 Integrated Space/Terrestrial Mobile Networks. He also contributed to the resource management activity of the Working Group 3000 within the RACE Project called Satellite Integration in the Future Mobile Network (SAINT, RACE 2117). From 1997 to 1998 he was with OTE of the Marconi Group, Florence, Italy, where he was involved in a GSM development program. In the same period he also contributed to the COST 252 Project (Evolution of Satellite Personal Communications from Second to Future Generation Systems) research activities. Since 1999, he has been a research associate at the Information Engineering Department, University of Siena, Italy, where he is involved in the activities of the Personalised Access to Local Information and services for tOurists (PALIO) IST Project within the fifth Research Framework of the European Commission. His research interests include third-generation mobile communication systems, medium access control protocols, traffic scheduling algorithms, and queuing theory.

BIBLIOGRAPHY

- 1. F. Harvey, The Internet in your hand, Sci. Am. (Oct. 2000).
- 2. K. J. Bannan, The promise and perils of WAP, Sci. Am. (Oct. 2000).
- 3. WAP Forum Website with address: http://www.wapforum.org/.
- B. Hu, Wireless portal technology an overview and perspective, Proc. 1st First On Line Symp. Electrical Engineers, Oct. 2000.
- WAP Forum, Wireless Application Protocol, WAP 2.0 Technical White Paper, available at the WAP forum site, http://www.wapforum.org/, Aug. 2001.
- 6. Example of WAP gateway characteristics: Nokia WAP Gateway, available at the address (date of access: Dec. 2001), http://www.nokia.com/corporate/wap/gateway.html.
- Wireless Application Protocol Forum, Ltd., WAP-154, Binary XML Content Format Specification, version 1.2, Nov. 4, 1999.
- 8. S. Lee and N.-O. Song, Experimental WAP (Wireless Application Protocol) traffic modeling on CDMA based mobile wireless

2906 WIRELESS ATM

network, Proc. 54th Vehicular Technology Conf., 2001, VTC 2001, 2001, pp. 2206–2210.

- 9. D. Ralph and H. Aghvami, Wireless application Protocol overview, Wireless Commun. Mobile Comput. J. 1(2): 125–140 (April–June 2001).
- 10. Wireless Application Protocol Forum Ltd., WAP-158, Wireless Datagram Protocol Specification, Nov. 5, 1999.
- A. Andreadis, G. Benelli, G. Giambene, and B. Marzucchi, Analysis of the WAP Protocol over SMS in GSM networks, Wireless Commun. Mobile Comput. J. 1(4): 381–395 (Oct.-Dec. 2001).
- 12. WAP browsers: Nokia, http://www.nokia.com Ericsson, http://www.ericsson.se/WAP UP.Browser from Phone.com, http://updev.phone.com WinWAP, http://www.slobtrot.com Motorola, http://www.slobtrot.com Gelon.net, http://www.gelon.net WAPman from Palm, http://palmsoftware.tucows.com.
- 3GPP, Technical Specification Group Terminals; Mobile Station Application Execution Environment (MExE); Functional Description; Stage 2. (3G TS 23.057).

WIRELESS ATM

NIKOS PASSAS LAZAROS MERAKOS University of Athens Panepistimiopolis, Athens, Greece

1. INTRODUCTION

Broadband and mobile communications are presently the two major drivers in the telecommunications industry. Asynchronous transfer mode (ATM) is considered the most suitable transport technique for the Broadband Integrated Services Digital Network (BISDN), because of its ability to flexibly support a wide range of services with quality-of-service (QoS) guarantees. These services are categorized in five classes according to their traffic generation rate pattern: constant bit rate (CBR), real-time variable bit rate (RTVBR), non-real-time variable bit rate (NRTVBR), available bit rate (ABR), and unspecified bit rate (UBR). On the other hand, wireless communications are enjoying a large growth in the last decade. Wireless local-area networks (LANs) in particular are becoming popular for indoor data communications because of their tetherless feature and increasing transmission speed. The combination of wireless communications and ATM, referred to as wireless ATM, aims at providing freedom of mobility with service advantages and QoS guarantees.

Wireless ATM is mainly considered for wireless access to a fixed ATM network; in this sense, it is applicable mostly to wireless LANs. A typical wireless ATM network (Fig. 1) includes the following main components:

• *Mobile terminals* (MTs), the end user equipment, which are basically ATM terminals with a radio adapter card for the air interface

- Access points (APs), the base stations of the cellular environment, which the MTs access to connect to the rest of the network
- An *ATM switch* (SW) to support interconnection with the rest of the ATM network
- A *control station* (CS), attached to the ATM switch, containing mobility specific software, to support mobility related operations, such as handover,¹ which are not supported by the ATM switch

In many proposals, the CS is considered integrated with the ATM switch in one network module, referred to as *switch workstation* (SWS). Even though this is the most common architecture, other schemes are possible. For example, APs could be equipped with switching and buffering capabilities, as proposed by Veeraraghavan et al. [1]. This, in principle, could expedite mobility and call control operations, but could also increase the overall cost of the system significantly, since the APs need to be more complicated, implementing the full signaling ATM stack.

The main challenge for wireless ATM is to harmonize the development of broadband wireless systems with BISDN/ATM, and offer similar advanced multimedia, multiservice features for the support of time-sensitive voice communications, LAN data traffic, video, and desktop multimedia applications to the wireless user. A sensible quality degradation is unavoidable, due to the reduced bandwidth of the wireless channel and the presentation capabilities of the MTs, but the network should be able to guarantee a minimum acceptable quality. Toward this direction, there are several problems to be faced, mainly because of the incompatibilities of the ATM protocol and the wireless channel:

- 1. ATM was originally designed for reliable, point-topoint optical fiber links. On the contrary, the wireless channel is a multiple access channel that suffers from high, time-varying, bit error rates, mainly due to fading and interference. This leads to the need for advanced multiple access control and error control mechanisms, for the efficient and reliable sharing of the scarce available bandwidth of the wireless channel, among different kinds of connections.
- 2. ATM was also designed for large bandwidth environments, following a bandwidth consuming policy to attain simplicity and fast switching of data packets. This leads to a packet header (ATM cell header, in the ATM terminology), which consumes approximately 10% of the available bandwidth (5 of 53 bytes). For gigabit-per-second (Gbps) optical fibers used in BISDN, this is not considered a drawback, compared to fast switching and packet delivery. But for a wireless channel of tens of megabits per second (Mbps), this can be vital for the overall performance. As shown later in this article, the usual practice is to perform header compression to reduce overhead as much as possible.

 $^1\,{\rm Mobility}$ issues will be explained in detail later in this article.



Figure 1. A typical wireless ATM network.

3. ATM signaling enhancements are definitely an important subject for wireless ATM, mainly for mobility. To support it, several additional functions and signaling need to be added in traditional ATM, for registration, location update, handover, and other applications. Particularly for handover, the comparatively high transmission speed, combined with the requirements of some real-time applications (e.g., videoconference), ask for fast and efficient handover techniques.

All these mobility-related functions are usually implemented in the CS shown in Fig. 1 to leave the conventional ATM switches intact. Mobility issues are discussed in detail later in this article. Additionally, some standard call control procedures of fixed ATM need also to be enhanced to cover the particularities of the wireless channel. Especially connection setup requires advanced call admission control algorithms that consider the instabilities of the wireless channel.

The rest of the article is organized in two main sections. Section 2 describes the basic issues and solutions for the medium access control, concluding with the most important standards. Some important protocols are discussed, and their effectiveness in servicing ATM traffic is analyzed. In Section 3, the required signaling enhancements for call and mobility control are discussed. The section starts with a basic signaling architecture, and continues with connection setup (especially call admission control), and handover. Finally, Section 4 contains our conclusions.

2. MEDIUM ACCESS CONTROL (MAC)

2.1. MAC Protocol Structure

In wireless ATM networks, an advanced MAC protocol is required, able to provide adequate support to the traffic classes defined by ATM standards, together with efficient use of the scarce radio bandwidth. Additionally, this protocol should be adaptive to frequent variations of channel quality.

MAC protocols can be grouped, in general, into five classes [2]: (1) fixed assignment, (2) random access, (3) centrally controlled demand assignment, (4) demand assignment with distributed control, and (5) adaptive strategies. Fixed-assignment techniques permanently reserve one constant capacity subchannel for each connection for its whole duration and they perform very well with constant bit rate connections in terms of both service quality and channel efficiency. However, their performance decreases dramatically when they are asked to support many infrequent users with variable-rate connections. In such cases, random-access protocols usually perform better. A typical example of such a protocol is Aloha, which permits users to transmit at will; whenever a collision occurs, collided packets are retransmitted after some random delay. It is well known that, although ALOHA-type protocols are easy to implement and attain minimum delays under light load, they suffer from long delays and instability under heavy traffic load. Enhancements of ALOHA include collision resolution techniques that increase the maximum achievable stable throughput. Centrally controlled demand assignment protocols reserve a variable portion of bandwidth for each connection, adjustable to its needs. Unlike random-access techniques, these protocols operate in two phases: reservation and transmission. In the reservation phase, the user requests from the system the portion of bandwidth required for its transmission needs, and the system responds by reserving the bandwidth and informing the user, while in the second phase the actual transmission takes place. Demand assignment protocols are usually complex, but are also stable and perform well under a wide range of conditions, although the reservation phase results in time and bandwidth consumption. With distributed control, the users themselves schedule their transmissions, based on broadcast information. Finally, adaptive schemes combine elements from techniques 1-4, and aim at supporting many different types of traffic [3].

Concerning the multiple access technique, the proposed protocols for the radio interface of wireless ATM networks are in general based on frequency-division multiple access (FDMA), code-division multiple access (CDMA), or timedivision multiple access (TDMA), or combinations of these techniques. The scarcity of available frequencies, and the requirement for dynamic bandwidth allocation, especially for VBR connections render the use of FDMA inefficient. On the other hand, CDMA limits the peak bit rate of a connection to a relatively low value, which is a problem for broadband applications (>2 Mbps). Accordingly, most of the proposed protocols use an adaptive TDMA scheme, due to its ability to flexibly accommodate a connection's bit rate needs, by allocating a variable number of time slots, depending on current traffic conditions.

Beyond this general choice of a TDMA-based scheme, the MAC protocols proposed in the literature differ in the technique used to build the required adaptivity in the TDMA scheme. The three main techniques used, alone or in combinations, are *contention*, *reservation*, and *polling*.

Contention-based random-access protocols are simple and require minimal scheduling. An example is the slotted ALOHA with exponential backoff protocol presented by Porter and Hopper [4]. Functionality that can be omitted from the MAC layer, such as handover and wireless call admission control, is pushed to the upper layers. These protocols, attain good delay performance under light traffic, and fit well with the statistical multiplexing philosophy of ATM. Nevertheless, their performance is questionable under heavy traffic conditions, or when multiple traffic classes must be supported with guaranteed QoS.

Another group of protocols uses reservation techniques, mainly through reservation/allocation cycles, to dynamically allocate the available bandwidth to connections, based on their current needs and traffic load. A welldesigned representative protocol of this group can be found in the article by Raychaudhuri et al. [5]. It is a TDMA time-division duplex (TDD) protocol, where time is divided in constant length frames and every frame is subdivided in a request subframe and a data subframe. The request subframe is accessed by MTs, through a simple slotted-ALOHA protocol, in order to declare their transmission needs, while the data subframe is used for user data transmission. The allocation of data slots is performed by the AP, based on a scheduling algorithm, and the MTs are informed through broadcast messages. These protocols are more complex and introduce some extra delays, due to the required reservation phase; on the other hand, they are stable under a wide range of traffic loads and can guarantee a predictable quality of service, which is very important in wireless ATM networks. Their performance depends to a large extent on the scheduling mechanism used for the allocation of the available bandwidth. A number of scheduling algorithms has been proposed in the literature, which try to separate real-time and non-real-time connections. For example, a minimum bandwidth can be allocated to non-real-time connections, while real-time connections are served as soon as possible. A delay-oriented scheduling algorithm, referred to as prioritized regulated allocation delay oriented scheduling (PRADOS), has been proposed to meet the requirements of the various traffic classes defined by the ATM architecture [6]. In order for PRADOS to maximize the fraction of ATM cells that are transmitted before their deadlines, each ATM cell is initially scheduled for transmission as close to its deadline as possible. After that, a packetization process ensures that no time slots will be left empty.

A third group of protocols uses adaptive polling to distribute bandwidth among connections [e.g., 7]. A slot is given periodically to each connection, without request, based on its expected traffic. Compared to reservationbased protocols, these protocols are simpler, since there is no reservation phase, but their performance depends on the algorithm that determines the polling period for each connection. If the polling period is shorter than needed, then such protocols might suffer from low utilization, since many slots will be empty. On the other hand, if the polling period is longer than needed, they result in increased delays and poor QoS. The problem becomes more difficult for variable-bit-rate bursty connections. Several proposals suggest an adaptive algorithm to decide on the polling period of each connection, based on total traffic load, expected traffic for each connection, and required QoS [7].

Finally, to improve performance, a combination of the abovementioned schemes is possible; for example, a protocol that is based mainly on reservation, but has also a random-access part for urgent traffic. A typical representative of this category is mobile access scheme based on contention and reservation for ATM (MASCARA) [8]. The multiple access technique used in MASCARA for uplink (from the MTs to the AP of their cell) and downlink (from the AP to its MTs) is based on TDMA/TDD, where a time slot is equal to the time required to transmit an ATM cell. The MASCARA time frame is divided into a DOWN period for downlink data traffic, an UP period for uplink data traffic, and an uplink CONTENTION period used for MASCARA control information. Each of the three periods has a variable length, depending on the traffic to be carried on the wireless channel. The AP schedules the transmission of its uplink and downlink traffic and allocates bandwidth dynamically, based on traffic characteristics and QoS requirements, as well as the current bandwidth needs of all connections. The current needs of an uplink connection from a specific MT are sent to the AP through MT "reservation requests," which are either piggybacked in the data MPDUs (mobile power distribution units), where the MT sends in the UP period, or contained in special "control MPDUs" sent for that purpose in the CONTENTION period. Protocols belonging to the same category can be found in the literature [5,9].

To minimize overhead added by the ATM header, header compression techniques can be used. A straightforward solution is the replacement of the 3-byte-long VPI/VCI (virtual path identifier/virtual channel identifier), used for addressing in ATM, with a shorter MAC specific identifier (MAC_ID), whose length is at most 1 byte, depending on the environment. The MAC_ID is used only for wireless channel transmission, and after this it is replaced with the original VPI/VCI.

2.2. Error Control

In wireless ATM, fulfilling the strict QoS requirements of ATM over an unreliable wireless channel is a challenging problem, and error control is very important. The error control mechanisms used can be thought of as belonging to a sublayer of the MAC layer (usually the upper part), referred to as *wireless data-link control* (WDLC) sublayer. WDLC is responsible for recovering from occasional quality degradations of the wireless channel, and for providing an interface to the ATM layer in terms of frame format and required QoS.

Error control techniques, in general, can be divided in two main categories: <u>automatic repeat request</u> (ARQ) and forward error correction (FEC). In ARQ techniques, the receiver detects the erroneously received data and requests retransmission from the transmitter. Since retransmissions imply increased delays, ARQ is efficient for non-real-time data. ARQ techniques are conceptually simple and provide high system reliability at the expense of some extra delay and bandwidth consumption due to retransmissions. FEC, on the other hand, is efficient for real-time data. A number of bits is added in every transmitted data unit, using a predetermined errorcorrection code, which allows the receiver to detect and correct errors up to a predetermined number per data unit, without requesting any additional information from the transmitter. It is clear that FEC techniques are fast at the expense of lower bandwidth utilization because of the transmission of additional bits.

In wireless ATM, where both real-time and non-realtime data must be supported, a hybrid scheme combining ARQ and FEC is usually used. According to this, for real-time connections (e.g., CBR, RTVBR) FEC bits are included in the header of every MAC data unit, to allow the receiver (AP or MT) to correct most of the errors. For non-real-time connections (e.g., NRTVBR), no extra bits are included, and the AP (MT in the downlink) requests from the MT (AP in the downlink) the retransmission of erroneously transmitted MAC data units.

2.3. MAC Standards

Currently, the MAC technology for wireless ATM is served mainly by two standards, both based on TDMA. The 802.11 standard [10], developed by the IEEE 802 LAN standards organization, and the high-performance radio LAN type 2 (HIPERLAN/2) [11], defined by the European Telecommunications Standards Institute (ETSI) RES-10 Group. Although both standards were designed mainly for conventional LAN traffic, they can definitely serve as a medium for passing ATM traffic, with the proper QoS guarantees. Here we focus more on HIPERLAN/2 because it provides more flexibility for ATM traffic. IEEE 802.11 operates at 2.4 GHz and considers data traffic up to 2 Mbps. The medium can alternate between a contention mode, known as the contention period (CP), and a contention-free mode, based on polling, known as the contention-free period (CFP). IEEE 802.11 supports three different kinds of frames: management, control, and data. A management frame is used for MT association/deassociation, timing, synchronization, and authentication/deauthentication. A control frame is used for handshaking and positive acknowledgments during a CP, and to end a CFP. Finally, a data frame is used for transmission of data during a CP or CFP. On the horizon there is the need for higher data rates, for applications requiring wireless connectivity at 10 Mbps and higher. This will allow 802.11 to match the data rates of most wired LANs. There is no current definition of the characteristics for the higher data rate signal. However, for many of the options available to achieve it, there is a clear upgrade path for maintaining interoperability with 2-Mbps systems, while providing higher data rates as well.

HIPERLAN/2 systems, on the other hand, operate at the 5.2 GHz unlicensed band and attain transmission

rates ranging from 6 to 54 Mbps (a typical value is 25 Mbps). In that sense, it serves better the desired transmission speed for ATM applications. The MAC protocol of HIPERLAN/2 is based on a TDMA/TDD scheme. Time is divided in MAC frames, which are further divided into time slots. Time slots are allocated to the connections dynamically and adaptively depending on the current needs of each connection. Slot allocation is performed by a MAC scheduler that takes into account QoS requirements of each connection. A MAC scheduling algorithm has not vet been specified by the HIPERLAN/2 standards. An efficient algorithm that will be able to meet the requirements of different connections should be developed. The duration of each MAC frame is fixed to 2 ms. Each frame comprises transport channels for broadcast control, frame control, access control, downlink and uplink data transmission, and random access. All data between the AP and the MTs are transmitted in the dedicated time slots, except for the random access channel where contention for the same time slot is allowed. The length of the broadcast control field is fixed, while the length of the other field may vary according to the current traffic needs.

HIPERLAN/2 error control entity supports three different modes of operation: acknowledged mode, repetition mode, and unacknowledged mode. Acknowledged mode provides for reliable transmissions using retransmissions to compensate for the poor link quality. The retransmissions are based on acknowledgments from the receiver. The ARQ protocol that is used is selective-repeat (SR) allowing various transmission window sizes to be used depending on the requirements of each connection. In order to support QoS for delay critical applications (e.g., voice, real-time video), error control may also utilize a discard mechanism for discarding data units that have exceeded their lifetime. Repetition mode provides for reliable transmission by repeating data units. In repetition mode, the transmitter transmits new data units consecutively, and is allowed to make arbitrary repetitions of each data unit. No feedback is provided by the receiver. Finally, unacknowledged mode provides for unreliable, low-latency transmissions. In unacknowledged mode, data flow only from the transmitter to the receiver. No ARQ retransmission control or discard messages are supported.

From the above short description of the two standards, it is clear that HIPERLAN/2 provides more alternatives to better satisfy the requirements of different ATM connections. Nevertheless, we should note that, mainly as a result of increased complexity, HIPERLAN/2 products are not yet available in the market, while there is a wide range of 802.11 equipment from a number of vendors.

3. SIGNALING ENHANCEMENTS

Terminal mobility in wireless ATM requires a number of additional operations not supported in fixed ATM networks. These operations include the following:

Registration/Deregistration. When a MT is switched on, it needs to inform the network and be accepted by it to be able to send and receive calls. This operation is called *registration*. An important part of registration is authentication, where the MT is recognized as authentic and it is permitted to continue registering. The operation opposite to registration, when the MT is switched off, is called *deregistration*, and informs the network that the MT is no longer available.

- Location Update. When a MT has no active connections, it is practically untraceable by the network. So a passive operation is required, in which the system periodically records the current location of the MT in some database that it maintains, in order to be able to forward an incoming connection, when a new connection setup request arrives.
- Handover. (Also referred to as "handoff.") It is the operation that allows a connection in progress to continue as the MT changes channels in the same cell, or moves between cells. In a multichannel system, handovers within the same cell, where the connections are transferred to new radio channels, are referred to as *intracell handovers*. The case where the MT connections are transferred to an adjacent cell is referred to as an *intercell handover*. One of the key issues in wireless ATM is maintaining the QoS of different connections during a handover.
- Connection Setup. Standard protocols for connection setup in fixed ATM networks assume that the terminal's address implicitly identifies its attachment point to the network. However, this is not the case in wireless ATM. Thus, the ATM connection setup protocols must be augmented to dynamically resolve a MT endpoint location. Additionally, connection admission control (CAC), as part of the connection setup process, is much more difficult in wireless ATM. This is because the wireless channel quality varies in time, due to temporary interference or fading, so the available resources are not fixed. A proposal for wireless ATM CAC can be found in Yu and Leung [12].

Registration/deregistration and location update solutions are more or less generic and do not have extra requirements in a wireless ATM environment. Below we elaborate on connection setup and handover, and analyze their requirements and constraints, starting with the signaling architecture.

3.1. Signaling Architecture

Current trends in designing the access network (AN) part of fixed BISDN aim at concentrating the traffic of a number of different user-network interfaces (UNIs) and routing this traffic to the appropriate service node (SN) through a broadband V interface (referred to as VB). The main objective in AN design is to provide cost-effective implementations without degrading the agreed QoS, while achieving high utilization of network resources. This is reflected in both the reduction of the AN physical equipment and in the limitations imposed on the AN functionality, such as the inability to interpret the full ATM layer control information and signaling. The use of only low-level operations in AN forces the establishment of several internal mechanisms that are used to unambiguously identify the connection an ATM packet belongs to, and to convey only those connection parameters that are absolutely necessary for traffic handling.

In this framework, a fast control protocol running over a universal VB interface can be introduced [13], which serves a number of AN internal functions while preserving the highest possible degree of transparency at the SN. The protocol is based on the local exchange access network interaction protocol (LAIP), which was developed to accommodate the SN-to-AN communication requirements, as identified in the early study and design of the dynamic $VB_{5.2}$ interface, namely, the interface between the fixed ATM AN and the SN. In the relevant standardization bodies, the presence of such a protocol has been firmly decided and has been given the name Broadband Bearer Channel Control Protocol (B-BCCP). The services of the $VB_{5.2}$ control protocol enable the dynamic AN operation by conveying the necessary connection-related parameters required for dynamic resource allocation, traffic policing, and routing in the AN, as well as information on the status of the AN before a new connection is accepted by the SN.

The signaling access architecture for wireless ATM considered here is an extension of the broadband V interface, where an enhanced version of the VB_{5.2} control protocol is used to enable the dynamic operation of the AN and to serve the AN internal functions. It is assumed that a mobility-enhanced version of the existing B-ISDN UNI call control (CC) signaling is employed to provide the basic call control function and to support the handoverrelated functions. In addition, pure ATM signaling access techniques, based on metasignaling, are adopted for the unique identification and control of signaling channels. These features allow us to minimize the changes required to the signaling infrastructure used in the wired network, and, in this respect, they can guarantee the integration of the wireless ATM access system with fixed B-ISDN. However, when striving for full integration, the mobilespecific requirements imposed by the radio access part need to be taken into account.

In today's wired ATM environment, the user-network interface is a fixed port that remains stationary throughout the lifetime of a connection. The current B-ISDN UNI protocol stack uses a single protocol over fixed pointto-point or point-to-multipoint interfaces. On the other hand, in wireless ATM, mobility causes the user access point to the wired network to change constantly, and the mobile terminal connections must be transferred from access point to access point, through a handover process. The support of the handover functionality assumes that the fixed network of the access part has the capability to dynamically set-up and release bearer connections during the call. A well-accepted methodology to support these features is the call and bearer separation at the UNI. The use of the extended $VB_{5.2}$ interface control protocol for wireless ATM access systems serves for the setup and reconfiguration of fixed bearer connections of the same call, supporting in this way the call and bearer control separation in the AN part.

On the basis of the terminology described above, the following types of signaling interaction for the communication of peer entities can be identified [14]:

- Mobile Call Control Signaling (MCCS). This includes an enhanced B-ISDN call control signaling protocol (denoted as Q.2931*), based on the ITU (International Telecommunication Union) recommendation Q.2931, for the setup, modification, and release of calls between the MT and the CS. The enhancements required in the current signaling standards are related to the support of the handover function (e.g., inclusion of handover-specific messages).
- *Mobility Management Signaling (MMS)*. This is responsible for the MT registration/authentication and tracking procedures.
- Bearer Channel Control Signaling (BCCS). This serves for providing the traffic parameters to the AP, and handles the establishment, modification/reconfiguration, and release of fixed ATM connections between the AP and the CS.
- *Radio Channel Control Signaling (RCCS)*. This deals with low-level signaling related to the radio interface consisting of messages between the MT and the AP (MAC and physical layer specific messages).

At the user plane, the MT has a typical ATM protocol stack on top of a radio-specific physical layer and a MAC layer. The AP acts as a simple interworking unit that extracts the encapsulated ATM cells from the MAC frame, and forwards them to the CS through a proper ATM virtual connection. The MAC functionality realized at the AP is based on a MAC scheduler, which, on the basis of the ATM connection characteristics declared at connection setup and current transmission requests, allocates the radio bandwidth according to the declared QoS requirements and service type of each connection. As already mentioned, such a mechanism provides a degree of transparency to a subset of broadband/ATM services, and achieves efficient sharing of the scarce radio bandwidth among the mobile users. The CS realizes the typical B-ISDN protocol functionality of the U plane.

3.2. Connection Setup

Connection setup procedures used in traditional ATM networks assume (1) reliable gigabit links with fixed capacity and (2) stationary users. Accordingly, CAC algorithms do not need to be constantly informed about the available resources and the users' attachment points. But this is not the case in wireless ATM. The wireless channel impairments and MAC layer overheads can result in lower bandwidth than the theoretically available, while a mobile users' attachment point with the network can change anytime. Below we describe typical connection setup scenarios in wireless ATM, focusing on the differences with fixed ATM.

When a MT initiates a new call, its signaling channel transparently conveys a standard connection SETUP_REQUEST signaling message to the CS. Upon receipt of this request, the CS identifies the calling MT and the called terminal, and contacts the location server to track the location of the calling MT and the called terminal (if it is mobile). An initial connection acceptance decision is made, based on the user service profile data and on the QoS requirements set by the MT.

In case the request is accepted, the AP of the calling MT should be notified by the CS (using BCCS) on the expected new traffic so that it can decide on the admission in the wireless channel, and allocate radio resources accordingly. To this end, the traffic parameters of the new connection, or at least a useful subset of them, should be communicated to the AP of the calling MT. This information makes it possible to exercise a policing functionality at the AP, implemented implicitly by its radio bandwidth allocator. It also protects the CS from the unlikely case where, although the CS expects availability of radio resources, these are exhausted due to additional overheads of the MAC layer, or a temporary reduction in radio link quality. The latter is useful in case the CAC of the CS does not take into account issues specific to the wireless access. Since the final CAC decision is taken at the CS, it is possible to implement a connection acceptance algorithm customized to the specific wireless access system. Traffic characteristics will appear at the AP together with the QoS requirements, declared as the class of service (CoS) that the specific connection will support. This enables the MAC to implement a set of priorities according to the connection to which an ATM cell belongs. To be able to recognize the particular connection class, it is necessary to declare also the VPI/VCI values that will be used.

The task of the AP-CS communication and bearer channel establishment in the fixed access network is very important in this case. An ALLOC message is generated and forwarded to the AP, through BCCS. The AP will reply with an ALLOC_COMPLETE or an ALLOC_REJECT message indicating whether it agrees with the CAC decision. The latter implies that the call is rejected at the AP. On receipt of an ALLOC_COMPLETE, the CS returns a CALL_PROCEEDING message to the calling MT and initiates the connection establishment procedures toward the core network (B-ISUP IAM message) if the called terminal is a fixed one.

In case the called terminal is another MT (i.e., intra-CS call), the call processing module of the CS forwards the setup request towards the AP of the called terminal, where functions similar to those described above take place. The signal exchanges for this case are shown in Fig. 2. In the fixed-to-MT (incoming) connection setup scenario, the CS receives an incoming SETUP_REQUEST message, identifies the called MT, tracks its location, draws an initial CAC decision, and asks the corresponding AP of the called MT [14].

In all cases, the ALLOC message transfers to the AP all the connection-related information required for the AP operation. This includes the bandwidth requested by the connection, the service class, the QoS parameter values, etc. An improvement, in the case where the requested bandwidth or the QoS cannot be supported by the radio part of the communication path, is for the AP to generate an ALLOC_MODIFY message indicating



Figure 2. Connection setup procedure between two MTs.

this situation and suggesting a QoS degradation needed for the connection to be accepted. This useful "fallback" mechanism intends to set up connections with the highest available bandwidth. However, such a capability is useless if the standard ATM signaling does not support QoS negotiation to let the CS and the MT negotiate the new situation. In all scenarios, we have implicitly assumed that MTs remain stationary at connection setup. If we assume that a MT may move during connection setup, the setup might not succeed. In this case, the new location of MT is determined and another setup should be attempted following the same procedures. The calling or called party can initiate the release of a call. On receipt of a RELEASE message, the CS releases all the resources associated with that call and triggers the release toward the AP, the core network, or the MT.

3.3. Handover

Among mobile-specific operations, handover is probably the most difficult to perform, due to the diversity of requirements of different kinds of connections, and the constraints imposed by the wireless channel. In any case, an unavoidable period of time is required, during which the end-to-end connection data path is incomplete. This means that some data might get lost or should be buffered for later delivery. The effect that this increase of losses or delays has on each application depends on the nature of the application and the duration of the disruption. Current proposed protocols for handover in wireless ATM may be grouped in four categories [15]:

- *Full-Connection Rerouting*. This is the simplest kind of handover, where the system establishes a completely new end-to-end route for each handover—as if it were a new connection. Clearly, this kind of handover is simple in terms of implementation, but can result in unacceptable delays and losses, depending on the distance between the two parties.
- Route Augmentation. In this case, the original connection is extended with an additional hop to the MT's next location. For users with limited mobility, this solution can result in low delays and very limited or no losses, since no actual rerouting is performed. But if the MT begins to change cells more often, the additional extensions will result in a very long connection path, increasing delays and reducing network utilization.
- Partial-Connection Rerouting. This kind of handover attempts to perform a more efficient rerouting, by preserving as much of the old connection path as possible and rerouting the rest. The key issue here is to locate the nearest ATM network node that is common to both the old and the new data paths. Then the common node will handle the tearing down of the old part and establishing the new, also taking care of the data that are on the way in the old part, when the switching is performed. Temporary buffering before switching or temporary rerouting after switching can be used to minimize losses. Partial connection rerouting is the most common handover type found in the literature.

Multicast Connection Rerouting. In this kind of handover, more than one connection paths are maintained at a time, although only one path is operational. When the MT moves to a new cell, data can immediately start flowing toward the new direction. This eliminates the need for establishing a new path during handover (partial or full) and leads to lower delays and losses. On the other hand, since the system cannot maintain a path for every cell a MT can move to, an intelligent algorithm is required to predict the MT's movement and preestablish paths on the neighboring cells, while at the same time, paths that are no longer needed are canceled. An extension of this kind of handover could also permit the same data to flow in more than one data path when the MT is at the threshold between two cells (this is also referred to as *macrodiversity*). This allows a MT with multiple receivers (antenna diversity) to get data from more than one AP, and keep only the correctly transmitted information, reducing in this way the bit error rate.

Another categorization in wireless ATM handover is based on who performs what. In general, a handover mechanism involves a continuous procedure of channel measurements, and starts with a handover request initiation. In that sense, there are three fundamentally different categories of handover mechanisms: network-controlled, mobileassisted, and mobile-controlled. In network-controlled handover, the MT is completely passive. All measurements are performed by the network (basically the AP) and the handover request is initiated by the AP. This is a simple solution, which does not perform well in the case where the signal received by the AP is good, while the signal received by the MT is bad. This weakness is overcome by mobileassisted handover, where both the AP and the MT are measuring the strength of the received signal; however, the handover request is initiated by the AP. The MT can only send its measurements to the BS in order for it to have a better picture of the situation. Finally, in mobile-controlled handover all measurements and handover requests are executed at the MT. If the handover request is executed via the "old" AP (the AP that the MT is leaving), we have a backward handover, and if it is executed via the "new" AP, we have a forward handover. Backward handovers are in general more seamless than forward, so the usual practice is for the MT to prefer backward handover, and, only if this is not possible (in case of an abrupt signal strength reduction), to perform forward. Mobile-controlled handover can operate either alone or in conjunction with network-controlled or mobile-assisted handover.

No matter what handover algorithm is used, the main target should be to maintain the QoS of active connections, not only during, but also after handover in the new cell. During handover, temporary buffering can be used at the switching point to ensure delivery of loss-sensitive data. For delay-sensitive data that cannot be buffered, the only solution is to ensure simple and fast handover operation. On the other hand, maintaining QoS in the new cell is not always possible. In fixed ATM, if an efficient CAC algorithm decides that a connection can be accepted with the requested QoS, then the network can guarantee this QoS throughout the duration of the connection. The same cannot be said for a connection to a MT, which can be rerouted when the MT is handed over to a new cell. For example, if this new cell is overcrowded, there might not be enough resources to support the QoS of the connections of the newly arrived MT. In this case, the smallest possible number of the MT's connections should be rejected, to leave enough resources for the rest. This decision is usually taken by the CS, because it has a more global view of the system. A more advanced solution is to renegotiate the QoS in the new cell in order to avoid connection rejection as much as possible. In this case, the MT will be asked to reduce its requirements if it wants to maintain its connections in the new cell. In the following paragraphs we describe a simple but typical mobile-controlled handover procedure.

When the MT decides that a handover should be performed, it sends a HANDOVER_REQUEST message toward the CS, transparently via the old AP. This message contains identification of the MT, the call, and the target AP. The MT may have multiple active connections at the same time, as multimedia applications are to be supported. If this is the case, during the request for handover the MT could also indicate the priorities of the different connections in case the new AP cannot accommodate all of them.

A fast control protocol between AP and CS is required for the release/establishment of the old/new bearers in the fixed network part, and for performing possible QoS renegotiations during handover. On receipt of the HANDOVER_REQUEST, the CS identifies the MT, and initiates a state machine for the handover. Similar procedures to those described for connection setup are performed between the CS and the new AP. In this way, the CS informs the target AP about the expected QoS and bandwidth requirements to allocate radio resources accordingly.

When the CS receives the response from the new AP (ALLOC_COMPLETE), it sends a HAN-DOVER_RESPONSE message to the MT to inform it about the handover results and possible QoS modifications, and reconfigures the ATM connections of the ATM switch toward the new AP. After receiving the HAN-DOVER_RESPONSE, the MT releases its radio connection with the old AP, and establishes a radio link with the new AP. Special ATM (and lower) layer cell relay functions take place at the MT and the CS to coordinate the switching of traffic, and to guarantee the transport of user data at an agreed QoS level in terms of cell loss, ordering, and delay. Finally, the CS updates the location server about the new location of the MT, and sends a RELEASE message to the old AP to notify it that the connection no longer exists and to de-allocate the corresponding radio resources.

The handover process described above is expected to be fast. In the unlikely case that a MT moves again before the handover is accomplished, handover is again attempted to the current destination AP, until it eventually succeeds. The forward handover scenario is similar to the backward one. The MT releases the old radio connection and communicates directly with the new AP. Since all signaling is passed through this new AP, a dynamic signaling channel allocation scheme is employed, in order for the MT to obtain a signaling channel for passing the messages to the CS.

4. CONCLUSIONS

The design of wireless ATM systems to offer ATM services to wireless users has attracted considerable attention during the past few years, and a large number of proposals exist in the literature dealing with specific design issues. The most important of these issues are the medium access control and the signaling enhancements.

Medium access control is much more demanding in wireless ATM than in traditional wireless networks, owing to the, often conflicting, requirements of the various ATM traffic types. The current trend is for flexible, TDMA/TDD protocols with variable time frame, enabled with a sophisticated traffic scheduling algorithm that adjusts the bandwidth given to a connection to its timevarying requirements, without violating the contract made with other active connections.

On the other hand, enhancements are required to standard ATM signaling, to cover issues such as wireless call admission control and handover. Wireless call admission control is part of the overall call admission control process, handling available resources in the wireless link. Since the available wireless bandwidth is time-varying, due to temporary deterioration of the radio signal, the procedure should always have up-todate information on the current status of the radio link. Finally, handover is a completely new issue for wireless ATM signaling. Handover mechanisms should be fast and efficient, in order to minimize losses or delays, which could influence the QoS provided to the user. The usual practice is to introduce a special-purpose control station, centrally located in the wireless ATM network, which handles all the extra signaling and implements wireless call admission control and handover mechanisms.

As a final comment we can say that the future trends in wireless communications tend to be toward wireless IP-based systems with QoS provision (i.e., IPv6), rather than wireless ATM. Nevertheless, the issues and problems are more or less the same, so techniques and mechanisms developed for wireless ATM can, with proper adjustments, be used in wireless IP as well.

BIOGRAPHIES

Nikos Passas received his B.S. degree in computer engineering in 1992 from the University of Patras, Patras, Greece, and a Ph.D. degree in computer engineering from the University of Athens, Athens, Greece, in 1997. In 1995, he joined the Greek National Research Center "Demokritos" as a network engineer, where he worked on network management. Since 1997, he has been a senior researcher in the Communication Networks Laboratory, where he has been working on wireless networks. His areas of interest are multiple access control, quality of service of wireless networks, and performance analysis of wireless communications. Lazaros Merakos received the Diploma in electrical and mechanical engineering from the National Technical University of Athens, Greece, in 1978, and his M.S. and Ph.D. degrees in electrical engineering from the State University of New York, Buffalo, in 1981 and 1984, respectively. From 1983 to 1986 he was on the faculty of the Department of Electrical Engineering and Computer Science at the University of Connecticut, Storrs, Connecticut. From 1986 to 1994 he was on the faculty of the Electrical and Computer Engineering Department at Northeastern University, Boston, Massachusetts Between 1993 and 1994 he served as director of the communications at the Digital Signal Processing Research Center at Northeastern University. In 1994, he joined the faculty of the University of Athens, Athens, Greece, where he is presently a professor in the Department of Informatics and Telecommunications, director of the Communication Networks Laboratory and the Networks Operations and Management Center. His research interests are in the design and performance analysis of broadband networks and wireless/mobile networks and services. He is the author of over 120 papers in the above areas. He was the recipient of the Guanella Award for the Best Paper presented at the 1994 International Zurich Seminar on Mobile Communications.

BIBLIOGRAPHY

- M. Veeraraghavan, M. J. Karol, and K. Y. Eng, Mobility and connection management in a wireless ATM LAN, *IEEE J. Select. Areas Commun.* 15(1): 50–68 (Jan. 1997).
- F. A. Tobagi, Multiaccess link control, in P. E. Green, Jr., ed., *Computer Network Architectures and Protocols*, Plenum Press, New York, 1982.
- E. Ayanoglu, K. Y. Eng, and M. J. Karol, Wireless ATM: Limits, challenges, and proposals, *IEEE Pers. Commun. Mag.* 3(4): 18-34 (Aug. 1996).
- J. Porter and A. Hopper, An ATM-Based Protocol for Wireless LANs, Olivetti Research Ltd. Technical Report 94.2, April 1994, available at ftp://ftp.cam-orl.co.uk/pub/docs/ORL/ tr.94.2.ps.Z.
- D. Raychaudhuri et al., WATMnet: A prototype wireless ATM system for multimedia personal communication, *IEEE J.* Select. Areas Commun. 15(1): 83–95 (Jan. 1997).
- N. Passas and L. Merakos, Traffic scheduling in wireless ATM networks, Proc. IEEE ATM '97 Workshop, Lisbon, Portugal, May 1997.
- C.-S. Chang, K.-C. Chen, M.-Y. You, and J.-F. Chang, Guaranteed quality-of-service wireless access to ATM networks, *IEEE J. Select. Areas Commun.* 15(1): 106–118 (Jan. 1997).
- 8. F. Bauchot et al., MASCARA: A MAC protocol for wireless ATM, *Proc. Int. Conf. Telecommunications*, Melbourne, Australia, April 1997.
- F. D. Priscoli, Adaptive parameter computation in a PRMA, TDD based medium access control for ATM wireless networks, *Proc. IEEE GLOBECOM'96*, London, Nov. 1996, Vol. 3, pp. 1779–1783.
- IEEE 802.11D5, Wireless LAN Medium Access Control (MAC) and PHysical Layer (PHY) Specifications, July 1996.

- 11. ETSI TR 101 683 (V1.1.1), Broad Radio Access Network (BRAN); HIgh Performance Radio Local Area Networks (HIPERLAN) Type 2; System Overview, Feb. 2000.
- O. T. W. Yu and V. C. M. Leung, Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN, *IEEE J. Select. Areas Commun.* 15(7): 1208–1225 (Sep. 1997).
- I. S. Venieris et al., Architectural and control aspects of the multi-host ATM subscriber loop, J. Network Syst. Manage. 5: 55–71 (1997).
- N. Loukas, N. Passas, L. Merakos, and I. Venieris, A signaling architecture for wireless ATM access networks, ACM Wireless Networks J. 6: 145–159 (Nov./Dec. 2000).
- I. F. Akyildiz et al., Mobility management in current and future communication networks, *IEEE Network* 12(4): 39–49 (July/Aug. 1998).

WIRELESS COMMUNICATIONS SYSTEM DESIGN

CHINTHA TELLAMBURA Monash University Clayton, Victoria, Australia

A. ANNAMALAI Virginia Tech Blacksburg, Virginia

1. INTRODUCTION

The wireless communications industry has been experiencing phenomenal annual growth rates exceeding 50% since the late 1990s. This degree of growth reflects the tremendous demand for commercial untethered communications services such as paging, analog and digital cellular telephony, and emerging personal communications services (PCS), including high-speed data, full-motion video, Internet access, on-demand medical imaging, real-time roadmaps, and anytime, anywhere videoconferencing. By 2002 subscriber rates for personal wireless services are expected to reach 70% of the population for industrial nations, and by 2004 these rates are expected to reach 17% of the population worldwide. Of these subscribers, it is expected that by 2005 half will have data-capable handsets, creating an even greater demand for wireless data services. Since wired broadband services such as digital subscriber loop (DSL) and cable modems have been slow to market, this will drive even more customers to wireless alternatives; by 2003 more than 34% of homes and 45% of businesses in the United States will be served by wireless broadband services. The first-generation cellular and cordless telephone networks, which were based on analog technology with frequency modulation, have been successfully deployed throughout the world since the early and mid-1980s. Second-generation (2G) wireless systems employ digital modulation and advanced call processing capabilities. Third-generation (3G) wireless systems will evolve from mature 2G networks, with the aim of providing universal access and global roaming. Introduction of wideband packet data services for wireless Internet up to 2 Mbps (megabits per second) will probably be the main attribute of 3G systems.

To meet this increasing demand, new wireless techniques and architectures must be developed to maximize capacity and quality of service (QoS) without a large penalty in the implementation complexity or cost. This provides many new challenges to system designers, one of which is ensuring the integrity of the data is maintained during transmission. The largest obstacle facing designers of wireless communications systems is the nature of the propagation channel. The wireless channel is nonstationary and typically very noisy as a result of fading and interference. The sources of interference could be natural (e.g., thermal noise in the receiver) or synthetic (human-made; e.g., hostile jammer, overlay communication), while the most common type of fading is caused by multipath effects, in which multiple copies of a signal arrive out of phase at the receiver and destructively interfere with the desired signal. Another problem imposed by multipath effects is delay spread, in which the multiple copies of a signal arriving at different times spread out each data symbol in time. The stretched-out data symbols will interfere with the symbols that follow, causing intersymbol interference. All of these effects can significantly degrade the performance and QoS of a wireless system. Another critical issue in wireless system development is channel capacity. The Shannon channel capacity may be conveniently expressed in terms of the channel characteristics as

$$C = B \log_2(1 + \gamma |H|^2) \tag{1}$$

where γ is the signal-to-noise ratio (SNR), *B* denotes the channel bandwidth, and $|H|^2$ is the normalized channel power transfer characteristic. The ratio *C/B*, called *spectral efficiency*, is the information rate per hertz, is directly related to the modulation of a signal. To illustrate this, the analog AMPS cellular telephone system has a spectral efficiency of 0.33 bps/Hz while the digital GSM system has a spectral efficiency of 1.35 bps/Hz and the IS 54 system has 1.6 bps/Hz.

To overcome the problems mentioned above and to increase spectral efficiency, many techniques are employed, including

- The use of a set of signals that fade independently is referred to as diversity combining. Diversity techniques include selective combining, switched combining, maximal ratio combining, and equal gain combining. The effectiveness of diversity combining is limited by the degree of independence of fading within the set of signals. A measure of this can be obtained from calculating correlations between pairs of signals.
- Diversity can also be sought through the use of coding techniques, multiple frequency bands, and multiple antennas.
- "Smart" or directional antennas allow the energy transmitted toward the significant scatterers to be reduced and hence reduce far-out echoes.

2916 WIRELESS COMMUNICATIONS SYSTEM DESIGN

- Adaptive filters and equalizers can be used to flatten the channel response for wideband fading channels.
- The delay spread affects high-data-rate systems. The required data can be simultaneously transmitted on a large number of carriers, each with a low data rate, and the total data rate can be high. This concept, which is known as *orthogonal frequency-division multiplexing* (OFDM), is used in digital broadcasting.

The system designers need to assess the efficacy of such techniques to determine the most appropriate choice of complexity and implementation constraints. One may use Monte Carlo simulations or develop an analytic framework for system design. The analytic approach has three advantages over the Monte Carlo approach; it

- Facilitates rapid computation of the system performance
- Provides insight as to how different design parameters affect the overall system performance
- Provides some ability to optimize the design parameters

Nevertheless, analytic solutions are governed by a set of simplifying assumptions needed for analytic tractability, and hence care must be exercised when extrapolating from analytic results to real-world designs. However, as this approach can identify viable design options before further computer simulations are undertaken, it can be the first step of the design process of communication systems.

1.1. Fading Channels

A generic communication system is shown in Fig. 1. The information from the source is converted into a signal suitable for sending by the transmitter and is then sent over the channel. The channel is a description of how the communications medium alters the signal that is being transmitted. Finally the receiver takes the signals that have been altered by the channel, and attempts to recover the information that was sent by the source. The estimate of this information is passed to the sink as the received information. The channel modifies the signal in ways that may be unpredictable to the receiver, so the receiver must be designed on the basis of statistical principles to estimate the information and to deliver the information to the receiver with as few errors as possible. In our case, the channel is the wireless channel that includes all the antenna and propagation effects within it.

We now briefly describe the statistical models of fading multipath channels, which are frequently used in the analysis and design of wireless communications systems.

1.2. Fading Channel Characterization

In the most general setting, a fading multipath channel is characterized as a linear, time-varying system having an (equivalent lowpass) impulse response $h(t, \tau)$ (or a time-varying frequency response) H(t, f), which is a widesense stationary random process. Time variations in $h(t, \tau)$ or H(t, f) result in frequency spreading, which is known as *Doppler spreading*, of the signal transmitted through the channel. Multipath propagation results in spreading the transmitted signal in time. Consequently, a fading multipath channel may be generally characterized as a doubly spread channel in time and frequency.

The channel output y at time t can be found from the convolution of the input signal x(t) with the impulse response $h(t, \tau)$ (also known as the *input delay spread function*) of the channel at time t. We then have

$$y(t) = \int_{-\infty}^{\infty} h(t,\tau) x(t-\tau) \, d\tau \tag{2}$$

where τ is the delay variable.

Assuming that the multipath signals propagating through the channel at different delays are uncorrelated, we can characterize a doubly spread channel by the delay Doppler spread function, which is obtained by transforming $h(t, \tau)$ with respect to time. The scattering function $S(\tau, \nu)$ is a measure of the power spectrum of the channel at delay τ and frequency offset ν (relative to the carrier frequency). From the scattering function, we obtain the delay power spectrum of the channel (also called the *multipath intensity profile*) by simply averaging over

$$S_c(\tau) = \int_{-\infty}^{\infty} S(\tau, \nu) \, d\nu \tag{3}$$

This spectrum expresses the average power received for a transmitted pulse as a function of time delay, τ . The range of values over which the delay power spectrum $S_c(\tau)$ is nonzero is defined as the multipath spread of the channel T_m . Similarly, the Doppler power spectrum is

$$S_c(\nu) = \int_{-\infty}^{\infty} S(\tau, \nu) \, d\tau \tag{4}$$

This spectrum expresses the average power received for a transmitted pulse as a function of frequency offset, ν .

1.2.1. Doppler Spectrum. When a mobile moves at a certain velocity, as pathlengths between transmitter and receiver change, the Doppler effect results in a change of the apparent frequency of the arriving wave. The amount of this change is known as the Doppler shift. The maximum Doppler shift f_d is given by

$$f_d = \frac{vf_c}{c}$$



Figure 1. Generic communication model.

where v is the velocity of the mobile, f_c is the communication frequency, and c is the velocity of propagation of light.

Example 1. A mobile system operates at 900 MHz. What is the maximum Doppler shift observed by a mobile traveling at 80 km/h?

The maximum Doppler shift is

$$f_d = f_c rac{v}{c} = 900 imes 10^6 imes rac{80 imes 10^3}{60 imes 60 imes 3 imes 10^8} = 67 \; {
m Hz}$$

With multipath propagation, the copies of a signal arrive from several directions and each copy has its own Doppler frequency. Thus, the exact shape of the resulting spectrum $S_c(\nu)$ depends on the relative amplitudes and directions of each of the incoming signals. The range of values over which the $S_c(\nu)$ is nonzero is defined as the Doppler spread f_d of the channel. The exact expression for $S_c(\nu)$ cannot be obtained without making some assumptions of the arrival angle of the multipath signals. Most commonly, the arriving multipath signals at the mobile are assumed to be equally likely to come from any horizontal angle. The classic Doppler spectrum is then given by

$$S_c(\nu) = \frac{1.5}{\pi f_d \sqrt{1 - (\nu/f_d)^2}} \quad \text{for } |\nu| < f_d \tag{5}$$

and $S_c(v) = 0$ for $|v| \ge f_d$. This function is sharply limited to $\pm f_d$. The width of the Doppler spectrum is known as the *fading bandwidth*. This function (Fig. 2) is used as the basis of many simulators of mobile radio channels. The assumption of uniform angle-of-arrival distribution may not hold over short distances where propagation is dominated by the effect of a particular local scatterers, but this is a good reference model for the long-term average Doppler spectrum.

The Doppler power spectrum cannot be measured accurately in practice. The level crossing rate $\left(LCR\right)$ and

6

5

4

3

2

1

0

-1

-0.6

-0.8

10 $\log_{10} [\pi f_d S_c(f)]$

the average fade duration (AFD) are directly related to a given Doppler spectrum. These parameters are easier to measure. The LCR is the number of positive-going crossings of a reference level r in unit time and the AFD is the average time between negative and positive level crossings. For the classic Doppler spectrum, the LCR is given by

$$N_r = \sqrt{2\pi} f_d \overline{r} e^{-\overline{r}^2} \tag{6}$$

where $\overline{r} = r/r_{\rm rms}$. The average fade duration for a signal level of \overline{r} is given by

$$\tau_{\overline{r}} = \frac{e^{-\overline{r}^2} - 1}{\sqrt{2\pi} f_d \overline{r}} \tag{7}$$

These two parameters are plotted in Figs. 3 and 4. Note that the signal spends most of its time crossing signal levels just below $r_{\rm rms}$, and that fades below this level have short duration.

1.2.2. Signal Correlation. The Doppler spread provides a measure of how rapidly the channel impulse response varies in time. The inverse Fourier transform of the Doppler power spectrum (5) is the autocorrelation function (ACF), which expresses the correlation between a signal at time t and $t + \tau$. For a classic spectrum (5) with Rayleigh fading, the correlation function is

$$\rho(\tau) = J_0(2\pi f_d \tau) \tag{8}$$

where $J_0(x)$ is the Bessel function of the first kind and zeroth order. This is plotted in Fig. 5. For large f_d , the correlation can decrease rapidly. To express this temporal relationship, the channel coherence time T_c for a channel is defined as the time over which the channel can be assumed constant. This is assured if the ACF remains close to unity for this duration. The coherence time is therefore inversely proportional to the Doppler spread of the channel:

$$T_c \propto \frac{1}{f_d} \tag{9}$$



Figure 2. The classic Doppler spectrum.

Figure 3. The normalized LCR for the classic Doppler spectrum.





Figure 4. The AFD for the classic Doppler spectrum.



Figure 5. ACF for the classic Doppler spectrum.

Thus a slowly fading channel has a large coherence time, and a rapidly fading channel has a small coherence time. To determine the proportionality constant, a threshold level of correlation for the complex envelope has to be chosen. A useful approximation to the coherence time for the classic channel is

$$T_c \approx \frac{9}{16\pi f_d} \tag{10}$$

Example 2. A mobile system operates at 900 MHz. The maximum speed of a mobile is 80 km/h. What is the minimum symbol rate to avoid the effects of Doppler spread?

From the previous example, the maximum Doppler shift is 67 Hz. The coherence time is therefore

$$T_c pprox rac{9}{16\pi imes 67} = 2.7 ext{ ms}$$

So the minimum symbol rate for undistorted symbols is, the reciprocal of this, 500 symbols per second. As most systems have data rates exceeding this, the correlation effect is negligible on most practical systems.

The channel coherence bandwidth is defined as the reciprocal of the multipath spread

$$B_c = \frac{1}{T_m} \tag{11}$$

If the correlation is examined for two signals at the same time, then the frequency separation for which the correlation equals 0.5 is termed the *coherence bandwidth* of the channel. This measures the width of the band of frequencies that are similarly affected by the channel response, that is, the width of the frequency band over which the fading is highly correlated.

The product $T_m f_d$ is called the spread factor of the channel. A spread factor smaller than unity results in an underspread channel. A spread factor greater than unity results in an overspread channel. For severely underspread channels $(T_m f_d \ll 1), h(t, \tau)$ can be measured by the use of nondata symbols. These symbols (pilot symbols) are known to the receiver and are regularly spread in time and/or frequency domains. Pilot symbols constitute an overhead. Channel measurements can be used at the receiver to demodulate the received signal and at the transmitter to optimize the transmitted signal.

1.3. Flat Fading

We define the time-varying transfer function of the channel as

$$H(f,t) = \int h(t,\tau) e^{-j2\pi f\tau} d\tau$$
(12)

Using this, the channel output (2) for a band-limited input signal x(t) can be expressed as

$$y(t) = \int_{f \in f_x} H(f, t) X(f) e^{j2\pi f t} df$$
(13)

where f_x denotes the frequency range over which X(f) is not zero [for $f \notin f_x$, X(f) = 0]. If the bandwidth of the signal is much less than the coherence bandwidth of the channel, H(f, t) does not change appreciably over the integration interval above. In other words, all the frequency components of x(t) are subject to the same attenuation and phase shift in transmission through the channel. Such a channel is called *frequency-nonselective*, *narrowband*, or *flat fading*. The effect of the channel on the signal is thus multiplicative and the channel output can be written as

$$y(t) = \alpha(t)x(t) \tag{14}$$

where $\alpha(t)$ is the complex fading coefficient at time *t*.

A frequency-nonselective channel is slowly fading if the time duration of a transmitted symbol T_s is much smaller than the coherence time of the channel ($T_s \ll T_c$). Equivalently, $T_s \ll 1/f_d$ or $f_d \ll 1/T_s$. A slowly fading, frequency-nonselective channel is normally underspread.

Table 1. Channel Types

	••		
	Flat	Selective	
Slow	$T_s < 1/f_d \ T_m < T_s$	$T_s < 1/f_d$ $T_m > T_s$	
Fast	$T_s > 1/f_d \ T_m < T_s$	$\begin{array}{l} T_s > 1/f_d \\ T_m > T_s \end{array}$	

A rapidly fading channel is defined by the condition $T_s \ge T_c$. Table 1 shows several channel types.

Example 3. In the GSM mobile cellular system, which operates at around 900 MHz, data are sent in bursts of duration approximately 0.5 ms. The maximum speed of a mobile is 80 km/hr. Is this a rapidly or slowly fading channel? The TETRA digital private mobile radio system, which operates at around 400 MHz, with a burst duration of around 14 ms. Is this a rapidly or slowly fading channel?

For the GSM case, $T_s f_d = 0.5 \times 10^{-3} \times 64 = 0.034$. For the TETRA case, the maximum Doppler is 40 Hz. So $T_s f_d = 14 \times 10^{-3} \times 40 = 0.5$.

1.4. Frequency-Selective Fading

When considering mobile/wireless systems for voice and low-bit-rate data applications, it is customary to use narrowband channels. But the wideband mobile radio channel has assumed increasing importance as the emergence of data rates to support multimedia services.

When the transmitted signal has a bandwidth greater than the coherence bandwidth of the channel, the signal suffers *frequency-selective* fading. Such channels also include time-selective fading. The standard model for wideband channel models is a tapped-delay line with complex-valued, time-varying tap gains. In the most general model, we have

$$h(t,\tau) = \sum_{n} \alpha_n(t)\delta(t-\tau_n(t))$$
(15)

The tap gains $\alpha_n(t)$ are usually modeled as stationary mutually uncorrelated random processes having not necessarily identical ACFs and Doppler power spectra. Thus each resolvable multipath component may be modeled with its own appropriate Doppler power spectrum (5) and corresponding Doppler spread.

1.5. Fading Distribution Models

A transmitter and receiver are surrounded by objects that reflect and scatter signals. For a large number of such objects, we can apply the central limit theorem to model $h(t, \tau)$ as a Gaussian random process. If this process has a mean of zero, the envelope of the channel impulse response at time *t* has a Rayleigh probability distribution and the phase is uniformly distributed; that is, the envelope

$$R = |h(t,\tau)| \tag{16}$$

has the probability density function (PDF)

$$f(r) = \frac{2r}{\Omega} e^{-r^2/\Omega}, \quad r \ge 0$$
(17)

where $\Omega = E(r^2)$. The Rayleigh distribution is characterized by this single parameter. For the frequencynonselective channel, the envelope is simply the magnitude of the channel multiplicative gain [Eq. (14)]. For the frequency-selective channel model, each of the tap gains $\alpha_n(t)$ [Eq. (15)] has a magnitude that can be modeled as Rayleigh fading.

1.5.1. Nakagami m Distribution. The Nakagami distribution (m distribution) is a versatile statistical distribution that can accurately model a variety of fading environments. It has greater flexibility in matching some empirical data than do the Rayleigh, lognormal, or Rice distributions owing to its characterization of the received signal as the sum of vectors with random moduli and random phases. It also includes the Rayleigh and the one-sided Gaussian distributions as special cases. Moreover, the m distribution can closely approximate the Rice distribution. The PDF for this distribution is [10]

$$f(r) = \frac{2}{\Gamma(m)} \left(\frac{m}{\Omega}\right)^m r^{2m-1} e^{-mr^2/\Omega}, \quad r \ge 0$$
(18)

where the parameter *m* is defined as the ratio of moments, called the *fading figure*:

$$m = \frac{\Omega^2}{E(R^2 - \Omega)^2)}, \quad m \ge \frac{1}{2} \tag{19}$$

1.5.2. Rice Distribution. The Rice distribution is used to characterize the signal in a line-of-sight (LoS) channel. The received signal consists of a multipath component, whose amplitude is described by the Rayleigh distribution, and a LoS component (also called the *specular component*) that has constant power. The PDF for the Rice distribution is

$$f(r) = \frac{r}{\sigma^2} e^{-(r^2 + s^2)/2\sigma^2} I_0\left(\frac{rs}{\sigma^2}\right)$$
(20)

where s^2 represents the power in the nonfading (specular) signal components and σ^2 is the variance of the corresponding zero-mean Gaussian components. If *s* is set to zero, this reduces to the Rayleigh PDF.

2. ANALYSIS TECHNIQUES

In this section, we illustrate several analysis techniques via examples. We will consider diversity reception, outage analysis, and trellis codes. We will show how to obtain theoretical expressions for parameters such as the error rate, the average output SNR, and the outage probability.

2.1. Diversity Reception

Diversity methods (implemented at either the receiver, the transmitter, or both) can be effective for combating the effects of multipath fading. Performance and complexity can be traded off against each other when implementing

2920 WIRELESS COMMUNICATIONS SYSTEM DESIGN

diversity techniques. For instance, consider the design of an antenna array receiver for millimeter-wave communications. Since the wavelength is less than 1 cm, several tens of array elements can be placed on the surface of a portable receiver. Classic signal combining techniques such as maximal-ratio combining (MRC), equal-gain combining (EGC), and selection combining (SC) may not be used with a large number of antenna elements (say, N) because of the need for N independent receivers, which is expensive and obeys the law of diminishing returns. An alternative is switched diversity combining (SDC), but the performance is worse. Thus, suboptimal receiver structures may exploit ordered statistics or a partitioned diversity combining scheme can be used to achieve the performance comparable to the optimal receiver but with considerably fewer electronics (hardware) and power consumption. While performance analysis of such schemes is beyond the scope of this article, the following techniques are a good starting point.

The basic premise of diversity is that the receiver processes multiple copies of the transmitted signal, where each copy is received through a distinct channel. If these channels are independent, then the chance of a deep fade occurring on all the channels simultaneously is small. Indeed, if a chance of a fade in a channel is *p*, the chance of a fade among N independent channels is p^N , which can be very small. This method requires N receiver circuits in the combiner. Each channel and the corresponding receiver circuit is called a branch. Two conditions are necessary for obtaining a high degree of improvement from a diversity combiner: (1) the fading in individual branches should have low cross-correlation—if the correlation is high, then deep fades in the branches can occur simultaneously, which negates diversity gain; and (2) the mean power from each branch should be almost equal.

2.2. Selection Combining

We will next show how selection combining can be analyzed for Rayleigh fading channels. The selection diversity combiner selects the branch that instantaneously has the highest SNR. The mathematical expression for the output SNR is simply

$$\gamma_{sc} = \max(\gamma_1, \gamma_2, \dots, \gamma_N) \tag{21}$$

where γ_i is the SNR for the *i*th branch. For Rayleigh fading, using Eq. (17), the probability that a branch having an SNR less that γ can be found as

$$\Pr(\text{SNR} < \gamma) = \left(1 - e^{-\gamma/\Gamma}\right) \tag{22}$$

If all the fading branches are independent, the probability of the output of the selection combiner having an SNR less than γ is just the abovementioned probability raised to the power *N*. Thus, we have

$$\Pr(\gamma_{sc} < \gamma) = \left(1 - e^{-\gamma/\Gamma}\right)^N \tag{23}$$

where Γ is the SNR at the input of each branch, assumed to be the same for all branches. If γ is very small compared

to the mean input SNR $\Gamma,$ we have

$$\Pr(\gamma_{sc} < \gamma) \approx \left(\frac{\gamma}{\Gamma}\right)^N$$
 (24)

Hence, the probability of a fade is simply the equivalent for a single-branch Rayleigh raised to the power N. Diversity gain is defined as the decrease in mean SNR to achieve a given probability of signal exceedance with and without diversity. The preceding shows that diversity gain increases with N. To further analyze performance, we need the probability density function (PDF) of the output. By differentiating (23) with respect to γ , we obtain

$$f_{\gamma_{sc}}(\gamma) = \frac{1}{\Gamma} \sum_{k=1}^{N} k(-1)^{k+1} \binom{N}{k} e^{-k\gamma/\Gamma}$$
(25)

This PDF can be used to derive expressions for various statistical parameters of the output. For example, the average output SNR is obtained as

$$\overline{\gamma}_{sc} = \int_0^\infty \gamma f_{\gamma_{sc}}(\gamma) \, d\gamma$$
$$= \Gamma \sum_{r=1}^N \frac{1}{r}$$
(26)

The bit error rate for optimum detection of binary phase shift keying (BPSK), differential phase shift keying (DPSK), coherent frequency shift keying (CFSK), and noncoherent frequency shift keying (NCFSK) in Gaussian noise can be given as

$$P_{e}(\gamma) = Q\left(\sqrt{2a\gamma}\right) \begin{cases} a = 1 & \text{BPSK} \\ a = \frac{1}{2} & \text{CFSK} \end{cases}$$

$$P_{e}(\gamma) = \frac{1}{2} \exp\left(-a\gamma\right) \begin{cases} a = 1 & \text{DPSK} \\ a = \frac{1}{2} & \text{NCFSK} \end{cases}$$

$$(27)$$

where γ is the instantaneous SNR and

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

We model the instantaneous SNR as a random variable with the PDF in (25). So the average output error rate is obtained by the formula

$$\overline{P}_e = \int_0^\infty P_e(\gamma) f_{sc}(\gamma) \, d\gamma \tag{28}$$

(29)

Thus, we obtain

$$\overline{P}_e = rac{1}{2}\sum_{k=1}^N (-1)^{k+1} {N \choose k} \left[1 - \sqrt{rac{a\Gamma}{a\Gamma+k}}
ight]$$

for BPSK and CFSK

$$\overline{P}_e = rac{1}{2}\sum_{k=1}^N k(-1)^{k+1} \binom{N}{k} rac{1}{k+a\Gamma}$$

for DPSK and NCFSK.

This averaging technique can be extended to other higherorder modulation schemes. We refer the reader to many papers on such topics.

2.2.1. Unequal Fading Branches. If the power in the fading branches is unequal (but independent), we need to modify the above analysis. Thus we can show that

$$\Pr(\gamma_{sc} < \gamma) = \prod_{k=1}^{N} \left(1 - e^{-\gamma/\Gamma_k} \right)$$
(30)

where Γ_k is the SNR at the *k*th input branch. Although equal branch powers (where Γ_k is constant) are needed to obtain maximum diversity benefit, better results are achievable in this case. The error rate performance of the above modulation methods can be derived similarly.

2.2.2. Dual-Branch SC Performance in Correlated Rayleigh Fading. The discussion above is premised on the assumption of independent fading. However, the branch signals in practical diversity systems can often be correlated. So the effects of correlation in fading among diversity branches on the error rates of digital receivers is of interest to the designers. Fairly comprehensive results have been developed for maximal-ratio combining (MRC), with arbitrary orders of diversity. The performance of MRC depends on the distribution of a sum $\sum \gamma_l$ of correlated signals, which is known for many cases. Unfortunately, performance analysis of selection diversity combiner in correlated fading is much more difficult.

For the dual-branch case with correlated Rayleigh fading, we can write the cumulative distribution function (CDF) of the SC output as

$$P(\gamma_{sc} \le \gamma) = 1 - \exp\left(-\frac{\gamma}{\Gamma}\right) \left[1 - Q(a, b) + Q(b, a)\right] \quad (31)$$

where Q(a, b) is the Marcum Q function, defined as

$$Q(a,b) = \int_b^\infty \exp\left(-\frac{a^2 + x^2}{2}\right) I_0(ax) \, dx$$

and

$$a = \sqrt{\frac{2\gamma\rho}{\Gamma(1-\rho)}}$$
 and $b = \sqrt{\frac{2\gamma}{\Gamma(1-\rho)}}$

where ρ is the normalized envelope covariance between the two branches. By differentiating (31) with respect to γ , we find

$$f_{\gamma_{sc}}(\gamma) = \frac{2}{\Gamma} \exp\left(-\frac{\gamma}{\Gamma}\right) \left[1 - Q(a, b)\right]$$
(32)

This PDF can be used to obtain performance statistics. For example, the average output SNR is obtained as

$$\overline{\gamma}_{sc} = \int_0^\infty \gamma f_{\gamma_{sc}}(\gamma) d\gamma$$

= $2\Gamma - \frac{\Gamma}{2}(1-\rho)^2$

$$\times \int_{0}^{2\pi} \frac{1}{(1 - \sqrt{\rho}\cos\theta)(1 + \rho - 2\sqrt{\rho}\cos\theta)} \frac{d\theta}{2\pi}$$
(33)
= $\Gamma \left[1 + \frac{1}{2}\sqrt{1 - \rho} \right].$

Note that for heavily correlated branches (e.g., $\rho \approx 1$), the average output SNR is simply the single-branch input SNR; that is, there is no diversity gain.

Using a technique similar to the derivation of (29), we can show that

$$\overline{P}_{e} = \frac{1}{2(1+a\Gamma)} \left(1 - \frac{a\Gamma(1-\rho)}{\sqrt{[2+a\Gamma(1-\rho)]^{2} - 4\rho}} \right)$$

for DPSK and NCFSK (34)

Again for heavily correlated branches (e.g., $\rho \approx 1$), the average output BER is simply that of the single-branch case; thus, there is no diversity gain.

2.3. Dual-Branch EGC Performance in Correlated Rayleigh Fading

EGC is of practical interest because it provides performance comparable to the optimal MRC technique but with greater simplicity. However, analyzing EGC receiver performance in fading is much more difficult. This is due to the difficulty of finding the PDF of the EGC output SNR, which depends on the square of a sum of N fading amplitudes. A closed-form solution to the PDF of this sum has been elusive for nearly 100 years (dating back to Lord Rayleigh), and indeed, even for the case of Rayleigh fading (mathematically simplest distribution), no solution exists for N > 2.

In an EGC combiner, the output of different diversity branches are first cophased and weighted equally before being summed to give the resultant output. The instantaneous SNR at the output of the EGC combiner is

$$\gamma_{\text{egc}} = \frac{\gamma_1 + \gamma_2 + 2\sqrt{\gamma_1\gamma_2}}{2}$$
$$= \frac{1}{2}(R_1 + R_2)^2$$
(35)

where γ_1 and γ_2 are the SNRs on individual branches and R_1 and R_2 denote to the signal amplitudes divided by the factor $\sqrt{2N_0}$ (i.e., normalized with respect the noise voltage).

We next show how the BER performance of EGC reception in correlated fading can be analyzed. For exact analysis of EGC, we need the characteristic function of $R_1 + R_2$ and therefore

$$\begin{split} \phi_{\gamma}(\omega) &= E\left\{e^{i\omega(R_{1}+R_{2})}\right\} \\ &= (1-\rho)e^{-\omega^{2}(1-\rho)\Gamma/4}\sum_{k=1}^{\infty}\rho^{k-1} \\ &\times \left[\frac{(2k-1)!}{2^{k-1}(k-1)!}D_{-2k}\left(-j\omega\sqrt{\frac{\Gamma(1-\rho)}{2}}\right)\right]^{2} (36) \end{split}$$

2922 WIRELESS COMMUNICATIONS SYSTEM DESIGN

where $D_p(z)$ is the parabolic cylinder function of order p. The average BER can be expressed as

$$\overline{P}_e = E\left\{Q\left(\sqrt{2a\gamma}\right)\right\} = E\left\{Q\left[\sqrt{a}(R_1 + R_2)\right]\right\}$$
(37)

where a = 1 for BPSK and $a = \frac{1}{2}$ for CFSK. Using an infinite series for the error function, we obtain

$$\overline{P}_e = \frac{1}{2} - \frac{2}{\pi} \sum_{\substack{n=1\\n \text{odd}}}^{\infty} \frac{\exp(-n^2 \omega_0^2/2)}{n} \operatorname{Im}\left[\phi_{\gamma}(n\omega_0 \sqrt{a})\right]$$
(38)

where ω_0 is a suitably small parameter. This is the exact solution for the performance of EGC reception in a correlated Rayleigh fading environment.

2.4. Mobile Outage Undershadowing

Consider computing the image function defined as

$$\phi_{\alpha}(s) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{e^{-x^2}}{1 + se^{\alpha x}} dx \tag{39}$$

where $s, \alpha > 0$. The Laplace transform of a Suzuki PDF is a special case with $\alpha = \sqrt{2}\sigma/4.34$, where σ is the standard deviation of shadowing in decibels. The range of interest may be $3 < \sigma \le 12$ and $0 < s \le 10^3$. This image function has extensive applications in evaluating the outage performance of multiuser mobile radio networks.

Using some analytic techniques (see listed references), we can show that

$$\phi_{\alpha}(s) = \frac{h}{\sqrt{\pi}} \sum_{n=-\infty}^{\infty} \frac{e^{-(nh-\ln s/\alpha)^2}}{1+e^{nh\alpha}} + E_c$$
(40)

where h is a small parameter controlling the correction term E_c . The value of h should not be too large or too small. It is found that a h value between 0.2 and 0.4 is sufficient for this application.

2.5. Outage Probability

Consider evaluating the probability of outage (outage) in a mobile fading environment. The instantaneous signal powers are modeled as random variables (RVs) \mathbf{p}_k , $k = 0, \ldots, L$, with mean \overline{p}_k . Subscript k = 0 denotes the desired signal and $k = 1, \ldots, L$ are for interfering signals. The outage is given by

$$\mathbf{P}_{\text{out}} = \Pr\left\{qI > \mathbf{p}_0\right\} \tag{41}$$

where $I = \mathbf{p}_1 + \cdots + \mathbf{p}_L$ and q is the power protection ratio, which is fixed by the type of modulation and transmission technique employed and the quality of service desired. Typically, 9 < q < 20 (dB). On introducing $\gamma = qI - \mathbf{p}_0$, we can readily find the moment-generating function (MGF) $\phi_{\gamma}(s)$.

Since the outage probability is $p_{\text{out}} = \Pr(\gamma < 0),$ we can show that

$$P_{\text{out}} = \frac{1}{2n} \sum_{i=1}^{n} \tilde{\phi} \left[\frac{(2i-1)\pi}{2n} \right] + R_n \tag{42}$$

where $\tilde{\phi}(\theta) = \operatorname{Re}\left[(1-j\tan(\theta/2))\phi_{\gamma}(c+jc\tan(\theta/2))\right]$ and the remainder term R_n vanishes rapidly. Although c can be anywhere between 0 and a_{\min} , the optimal location ensures that $|\phi_{\gamma}(c+j\omega)|$ decays as rapidly as possible for $|\omega| \to \infty$. This rapid decay occurs if s = c is the saddle point; thus, at s = c, $s^{-1}\phi_{\gamma}(s)$ achieves its minimum on the real axis. While this optimal c requires a numerical search, it is sufficient to use $c = a_{\min}/2$. This formula can be used to compute the outage probability for various mobile systems and fading channel configurations.

2.6. Trellis-Coded PSK

The performance of convolutional codes, Turbo codes, and trellis-coded modulation (TCM) schemes over wireless channels has received much attention. There are several methods to analyze the performance of such codes. Here we describe the evaluation of the union bound.

The union bound technique is based on

$$P_b \le \frac{1}{k} \sum_{\mathbf{z}, \hat{\mathbf{z}} \in \mathcal{C}} a(\mathbf{z} \to \hat{\mathbf{z}}) P(\mathbf{z} \to \hat{\mathbf{z}})$$
(43)

where k is the number of input bits per encoding interval, $P(\mathbf{z} \rightarrow \hat{\mathbf{z}})$ is the pairwise error probability (PEP), $a(\mathbf{z} \rightarrow \hat{\mathbf{z}})$ is the number of associated bit errors, and C is the set of all legitimate code sequences. But the evaluation of even the union bound is difficult since $P(\mathbf{z} \rightarrow \hat{\mathbf{z}})$ requires complex calculations. Thus, bounds on $P(\mathbf{z} \rightarrow \hat{\mathbf{z}})$ itself are used to compute (43), resulting in a weaker union bound. We next show a more general method to evaluate the union bound exactly, and this method is applicable to practical schemes such as differential detection and pilot-tone-aided detection. This approach can also be extended to schemes such as Turbo coding and spacetime codes.

2.6.1. System Model. The received complex sample at time n is

$$y_n = \alpha_n z_n + v_n$$

where α_n is the channel gain and v_n is an additive Gaussian noise sample. The following is used throughout the presentation:

- A1. z_n is a q-ary phase shift keying (PSK) symbol (i.e., $z_n \in \{e^{j2\pi k/q} \mid k = 0, 1, \dots, q-1\}$ and $j = \sqrt{-1}$).
- A2. Each α_n is a zero-mean, complex, Gaussian random variable (RV). The α_n terms are independent (i.e., ideal interleaving/deinterleaving) and identically distributed RVs.
- A3. Each α_n remains constant during a symbol interval (i.e., nonselective slow fading).
- A4. The receiver has some form of channel measurements given by $\hat{\alpha}_n$ that is a complex Gaussian RV. The correlation coefficient between α and $\hat{\alpha}_n$ is μ .

If $\mu = 1$, ideal channel measurements exist. For practical channel estimators $|\mu| \leq 1$. The more μ deviates from unity, the larger is the performance penalty.

2.6.2. Pairwise Error Event Probability. Consider two codewords $\mathbf{z} = \{z_1, z_2, \ldots, z_N\}$ and $\hat{\mathbf{z}} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_N\}$ of length *N*. The Viterbi decoder computes the path metrics and selects \mathbf{z} over $\hat{\mathbf{z}}$ according to the path metric difference *D*. The characteristic function of *D*, $\phi(j\omega) = E[e^{j\omega D}]$, can be written as

$$\phi(j\omega) = \prod_{n\in\eta} \frac{\Delta_n}{\omega^2 - j\omega + \Delta_n} \tag{44}$$

where $\Delta_n \stackrel{\Delta}{=} [1 + (1 - |\mu|^2)\gamma_s]/(|\mu|^2|z_n - \hat{z}_n|^2\gamma_s), \eta \stackrel{\Delta}{=} \{n \mid z_n \neq \hat{z}_n, n = 1, \dots, N\}$, and $\gamma_s = \overline{E_s}/N_0$ is the average signal-to-noise ratio. It can be shown that

$$P(\mathbf{z} \to \mathbf{\hat{z}}) = \frac{-1}{2\pi j} \int_{-\infty+j\varepsilon}^{\infty+j\varepsilon} \frac{\phi(j\omega)}{\omega} d\omega$$
(45)

where ε is a small positive number. An explicit expression for $P(\mathbf{z} \rightarrow \hat{\mathbf{z}})$, which cannot be used with the transfer function approach, can be obtained by solving for the residues of the contour integral. However, a transfer function can be defined with the *factors* of $\phi(j\omega)$.

2.6.3. Union Bound. Consider the evaluation of the union bound (43). Let $\mathbf{Z} = (Z_1, Z_2, ...)$ be a vector of formal variables. Define the generating function of the form

$$T(\mathbf{Z}, I) = \sum_{\mathbf{z}, \hat{\mathbf{z}} \in \mathcal{C}} I^{a(\mathbf{z} \to \hat{\mathbf{z}})} \prod_{n \in \eta} Z_n$$
(46)

where I is another formal variable. Moreover, let

$$D_n(\omega) \stackrel{\Delta}{=} \frac{\Delta_n}{\omega^2 - j\omega + \Delta_n} \tag{47}$$

The number of distinct values that $D_n(\omega)$ can take depends on the size of the signal constellation. The transfer function $T(\mathbf{D}(\omega), I)$ can be determined by a signal flow graph with the branch labels $I^v D_n(\omega)$ for uniform trellis codes. By contrast, for the union–Chernoff bound, the branches are labeled with $I^v(1 + 1/(4\Delta_n))^{-1}$.

Combining (43), (45) and (46), and using the standard analysis, it follows that

$$P_{b} \leq \frac{-1}{j2\pi k} \left. \frac{\partial}{\partial I} \left\{ \int_{-\infty+j\varepsilon}^{\infty+j\varepsilon} \frac{T(\mathbf{D}(\omega), I)}{\omega} \, d\omega \right\} \right|_{I=1}$$
(48)

where the partial derivative can be computed. While this integral has no analytical solution in general, its numerical computation poses little difficulty. Since $|T(\mathbf{D}(\omega), I)| \to 0$ as $|\omega| \to \infty$, simple techniques such as the Simpson method are adequate.

Example 4. Consider the performance of the twostate trellis-coded QPSK (Fig. 6) in Rayleigh fading. Using branch label gains, $D_n(\omega)$, the transfer function becomes

$$T(\mathbf{D}(\omega), I) = \frac{I\Delta_2\Delta_4}{(\omega^2 - j\omega + \Delta_4)(\omega^2 - j\omega + (1 - I)\Delta_2)}$$



Figure 6. State diagram.



Figure 7. Bit error performance of rate $\frac{1}{2}$ trellis-coded QPSK (differential detection) for fast Rayleigh fading.

where Δ_2 and Δ_4 are obtained with $|z_n - \hat{z}_n|^2$ equal to 2 and 4, respectively. Substituting this in (48), carrying out the integration, and evaluating the derivative at I = 1, one has

$$P_{b} \leq 1 - \frac{(1 + (1 - |\mu|^{2})\gamma_{s})}{2|\mu|^{2}\gamma_{s}} + \frac{3(1 + (1 - |\mu|^{2})\gamma_{s})^{2}}{4|\mu|^{4}\gamma_{s}^{2}} - |\mu|\sqrt{\frac{\gamma_{s}}{1 + \gamma_{s}}}.$$
(49)

This is the exact union bound for this TCM scheme. If differential detection is used in a flat fading land mobile channel, the correlation coefficient μ is given in Ref. 19, Eq. (32) as a function of the normalized maximum Doppler frequency f_dT . Figure 7 shows Eq. (49) for several f_dT values. Unless $f_dT = 0$, an error floor exists.

Example 5. Consider the trellis-coded 8-PSK scheme given in Ref. 20, Fig. 5. Its transfer function Ref. 20, Eq. (19) can be defined with the weight profiles obtained using $D_n(\omega)$. Consider pilot-tone-aided detection with μ given in Ref. 19, Eq. (40). Assume that the bandwidth of the pilot tone filter is $2f_d$ and the power-split ratio is $\sqrt{2f_dT}$. Figure 8 shows the simulation results and the union bound (48). At $P_b \approx 10^{-4}$, the simulation results are within 0.2, 1.0, and 1.5 dB of the union bound for f_dT of 0, 0.02, and 0.04, respectively.



Figure 8. Bit error performance of rate $\frac{2}{3}$, 4-state, trellis-coded 8PSK for fast Rayleigh fading and pilot-tone-aided detection.

BIOGRAPHIES

C. Tellambura received his B.Sc. degree with honors from the University of Moratuwa, Sri Lanka, in 1986, his M.Sc. in electronics from the King's College, UK, in 1988, and his Ph.D. in electrical engineering from the University of Victoria, Canada, in 1993. He was a postdoctoral research fellow with the University of Victoria and the University of Bradford. Currently, he is a senior lecturer at Monash University, Australia. He is an editor for the *IEEE Transactions on Communications* and the *IEEE Journal on Selected Areas in Communications* (Wireless Communications Series). His research interests include coding, communications theory, modulation, equalization, and wireless communications.

A. Annamalai received his B.E. degree in electrical and computer engineering in 1993 from University of Science of Malaysis, and his M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Victoria in 1997 and 1999, respectively. He was with Motorola Inc. as an RF design engineer from 1993 to 1995 and a postdoctoral research fellow at the University of Victoria in 1999. Since December 1999, he has been an assistant professor at the Virginia Polytechnic Institute and State University where he has been working on smart antennas and communication receiver designs. Dr. Annamalai has published over 60 papers in the field of wireless communications. He was the recipient of the 1997 Lieutenant Governor's medal, 1998 IEEE Daniel E. Noble Graduate Fellowship, 2000 NSERC Doctoral Prize, 2000 CAGS/UMI Doctoral Dissertation Award, and the 2001 IEEE Leon K. Kirchmayer Prize Paper Award for his work on diversity systems. He is an editor for the IEEE Transactions on Wireless Communications and Wiley's International Journal on Wireless Communications and Mobile Computing, an associate editor for the IEEE Communications Letters and is the technical program chair for VTC'2002 (Fall). His research interests are in high-speed data transmission on wireless links, smart antennas, multicarrier communications, mathematical modeling of radio channels, and wireless communications theory.

BIBLIOGRAPHY

- M. Zeng, A. Annamalai, and V. K. Bhargava, Harmonization of global third generation mobile systems, *IEEE Commun. Mag.* 38: 94-104 (Dec. 2000).
- M. Zeng, A. Annamalai, and V. K. Bhargava, Recent advances in cellular wireless communications, *IEEE Commun. Mag.* 37: 128–138 (Sept. 1999).
- 3. S. R. Saunders, Antennas and Propagation for Wireless Communication Systems, Wiley, 1999.
- E. Biglieri, J. Proakis, and S. Shamai, Fading channels: information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* 44: 2619-2692 (Oct. 1998).
- H. Hashemi, The indoor radio propagation channel, *IEEE* Proc. 81: 943-967 (July 1993).
- B. Sklar, Rayleigh fading channels in mobile digital communication systems. Part I: Characterization, *IEEE Commun.* Mag. 35: 136-146 (Sept. 1997).
- B. Sklar, Rayleigh fading channels in mobile digital communication systems. Part II: Mitigation, *IEEE Commun. Mag.* 35: 148-155 (Sept. 1997).
- 8. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill Series in Electrical Engineering: Communications and Signal Processing, McGraw-Hill, New York, 1995.
- W. C. Jakes, ed., Microwave Mobile Communications, Wiley, New York, 1974.
- M. Nakagami, The m-distribution, a general formula of intensity distribution of rapid fading, in W. G. Hoffman, ed., *Statistical Methods in Radio Wave Propagation*, Pergamon, Oxford, 1960.
- N. C. Beaulieu and A. A. Abu-Dayya, Analysis of equal gain diversity on Nakagami fading channels, *IEEE Trans. Commun.* 39: 225-234 (Feb. 1991).
- 12. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill, New York, 1966.
- N. C. Beaulieu, A simple series for personal computer computation of the error function Q(.), *IEEE Trans. Commun.* 37: 989–991 (Sept. 1989).
- A. Annamalai, C. Tellambura, and V. K. Bhargava, Equalgain diversity receiver performance in wireless channels, *IEEE Trans. Commun.* 48: 1732–1745 (Oct. 2000).
- J.-P. M. G. Linnartz, Exact analysis of the outage probability in multiple-user mobile radio, *IEEE Trans. Commun.* 40: 20-23 (Jan. 1992).
- J.-P. Linnartz, Narrowband Land-Mobile Radio Networks, Artech, House, Boston, 1993.
- A. Annamalai, C. Tellambura, and V. K. Bhargava, Simple and accurate methods for outage analysis in cellular mobile radio systems-a unified approach, *IEEE Trans. Commun.* 49: 303–316 (Feb. 2001).
- C. Tellambura, Evaluation of the exact union bound for trellis coded modulations over fading channels, *IEEE Trans. Commun.* 44: 1693-1699 (Dec. 1996).
- J. K. Cavers and P. Ho, Analysis of the error performance of trellis coded modulations in Rayleigh fading channels, *IEEE Trans. Commun.* 40: 74-83 (Jan. 1992).

 R. G. McKay, E. Biglieri, and P. J. McLane, Error bounds for Trellis-Coded MPSK on a fading mobile satellite channel, *IEEE Trans. Commun.* 39: 1750–1761 (Dec. 1991).

WIRELESS INFRARED COMMUNICATIONS

JEFFREY B. CARRUTHERS Boston University Boston, Massachusetts

1. INTRODUCTION

Wireless infrared communications refers to the use of freespace propagation of lightwaves in the near-infrared band as a transmission medium for communication [1-3], as shown in Fig. 1. The communication can be between one portable communication device and another or between a portable device and a tethered device, called an *access point* or *base station*. Typical portable devices include laptop computers, personal digital assistants, and portable telephones, while the base stations are usually connected to a computer with other networked connections. Although infrared light is usually used, other regions of the optical spectrum can be used (hence the term "wireless optical communications" instead of "wireless infrared communications" is sometimes used).

Wireless infrared communication systems can be characterized by the application for which they are designed or by the link type, as described below.

1.1. Applications

The primary commercial applications are as follows:

- Short-term cableless connectivity for information exchange (business cards, schedules, file sharing) between two users. The primary example is Infrared Data Association (IRDA) systems (see Section 4).
- Wireless local-area networks (WLANs) provide network connectivity inside buildings. This can either be an extension of existing LANs to facilitate mobility, or to establish ad hoc networks where there is no LAN. The primary example is the IEEE 802.11 standard (see Section 4).
- Building-to-building connections for high-speed network access or metropolitan- or campus-area networks.



Figure 1. A typical wireless infrared communication system.

• Wireless input and control devices, such as wireless mice, remote controls, wireless game controllers, and remote electronic keys.

1.2. Link Type

Another important way to characterize a wireless infrared communication system is by the "link type," which means the typical or required arrangement of receiver and transmitter. Figure 2 depicts the two most common configurations: the point-to-point system and the diffuse system.

The simplest link type is the point-to-point system. There, the transmitter and receiver must be pointed at each other to establish a link. The *line-of-sight* (LoS) path from the transmitter to the receiver must be clear of obstructions, and most of the transmitted light is *directed* toward the receiver. Hence, point-to-point systems are also called *directed LoS systems*. The links can be temporarily created for a data exchange session between two users, or established more permanently by aiming a mobile unit at a base station unit in the LAN replacement application.

In diffuse systems, the link is always maintained between any transmitter and any receiver in the same vicinity by reflecting or "bouncing" the transmitted information-bearing light off reflecting surfaces such as ceilings, walls, and furniture. Here, the transmitter and receiver are *nondirected*; the transmitter employs a wide transmit beam and the receiver has a wide field of view (FoV). Also, the LoS path is not required. Hence, diffuse systems are also called *nondirected non-LoS systems*. These systems are well suited to the wireless LAN application, freeing the user from knowing and aligning with the locations of the other communicating devices.

1.3. Fundamentals and Outline

Most wireless infrared communications systems can be modeled as having an output signal Y(t), and an input signal X(t), which are related by

$$Y(t) = X(t) \otimes c(t) + N(t) \tag{1}$$

where \otimes denotes convolution, c(t) is the impulse response of the channel and N(t) is additive noise. This article is organized around answering key questions concerning the system as represented by this model.

In Section 2, we consider questions of optical design. What range of wireless infrared communications systems does this model apply to? How does c(t) depend on the electrical and optical properties of the receiver and transmitter? How does c(t) depend on the location, size, and orientation of the receiver and transmitter? How do X(t) and Y(t) relate to optical processes? What wavelength is used for X(t)? What devices produce X(t) and Y(t)? What is the source of N(t)? Are there any safety considerations? In Section 3, we consider questions of communications design. How should a data symbol sequence be modulated onto the input signal X(t)? What detection mechanism is best for extracting the information about the data from the received signal Y(t)? How can one measure and improve the performance of the system? In Section 4, we consider



Figure 2. Common types of infrared communication systems: (a) point-to-point system; (b) diffuse system.

the design choices made by existing standards such as IRDA and IEEE 802.11. Finally, in Section 5, we consider how these systems can be improved in the future.

2. OPTICAL DESIGN

2.1. Modulation and Demodulation

What characteristic of the transmitted wave will be modulated to carry information from the transmitter to the receiver? Most communication systems are based on phase, amplitude, or frequency modulation, or some combination of these techniques. However, it is difficult to detect such a signal following nondirected propagation, and more expensive narrow-linewidth sources are required [2]. An effective solution is to use *intensity* modulation, where the transmitted signal's intensity or power is proportional to the modulating signal.

At the demodulator (usually referred to as a *detector* in optical systems), the modulation can be extracted by mixing the received signal with a carrier lightwave. This *coherent detection* technique is best when the signal phase can be maintained. However, this can be difficult to implement and additionally, in nondirected propagation, it is difficult to achieve the required mixing efficiency. Instead, one can use *direct detection* using a photodetector. The photodetector current is proportional to the received optical signal intensity, which for intensity modulation, is also the original modulating signal. Hence, most systems use intensity modulation with direct detection (IM/DD) to achieve optical modulation and demodulation.

In a free-space optical communication system, the detector is illuminated by sources of light energy other than the source. These can include ambient lighting sources, such as natural sunlight, fluorescent lamp light, and incandescent lamp light. These sources cause variation in the received photocurrent that is unrelated to the transmitted signal, resulting in an additive noise component at the receiver.

We can write the photocurrent at the receiver as

$$Y(t) = X(t) \otimes Rh(t) + N(t)$$

where R is the responsivity of the receiving photodiode [in amperes per watt (A/W)]. Note that the electrical

impulse response c(t) is simply *R* times the optical impulse response h(t). Depending on the situation, some authors use c(t) and some use h(t) as the impulse response.

2.2. Receivers and Transmitters

A transmitter or *source* converts an electrical signal to an optical signal. The two most appropriate types of device are the light-emitting diode (LED) and semiconductor laser diode (LD). LEDs have a naturally wide transmission pattern, and so are suited to nondirected links. Eye safety is much simpler to achieve for an LED than for a laser diode, which usually has very narrow transmit beams. The principal advantages of laser diodes are their high energy-conversion efficiency, their high modulation bandwidth, and their relatively narrow spectral width. Although laser diodes offer several advantages over LEDs that could be exploited, most short-range commercial systems currently use LEDs.

A receiver or *detector* converts optical power into electrical current by detecting the photon flux incident on the detector surface. Silicon p-i-n photodiodes are ideal for wireless infrared communications as they have good quantum efficiency in this band and are inexpensive [4]. Avalanche photodiodes are not used here since the dominant noise source is background light-induced shot noise rather than thermal circuit noise.

2.3. Transmission Wavelength and Noise

The most important factor to consider when choosing a transmission wavelength is the availability of effective, low-cost sources and detectors. The availability of LEDs and silicon photodiodes operating in the 800–1000-nm range is the primary reason for the use of this band. Another important consideration is the spectral distribution of the dominant noise source: background lighting.

The noise N(t) can be broken into four components: photon noise or shot noise, gain noise, receiver circuit or thermal noise, and periodic noise. Gain noise is only present in avalanche-type devices, so we will not consider it here.

Photon noise is the result of the discreteness of photon arrivals. It is due to background light sources, such as sunlight, fluorescent lamp light, and incandescent lamp light, as well as the signal-dependent source $X(t) \otimes c(t)$. Since the background light striking the photodetector is normally much stronger than the signal light, we can neglect the dependency of N(t) on X(t) and consider the photon noise to be additive white Gaussian noise with two-sided power spectral density $S(f) = qRP_n$ where q is the electron charge, R is the responsivity, and P_n is the optical power of the noise (background light).

Receiver noise is due to thermal effects in the receiver circuitry, and is particularly dependent on the type of preamplifier used. With careful circuit design, it can be made insignificant relative to the photon noise [5].

Periodic noise is the result of the variation of fluorescent lighting due to the method of driving the lamp using the ballast. This generates an extraneous periodic signal with a fundamental frequency of 44 kHz with significant harmonics to several megahertz. Mitigating the effect of periodic noise can be done using highpass filtering in combination with baseline restoration [6], or by careful selection of the modulation type, as discussed in Section 3.1.

2.4. Safety

There are two safety concerns when dealing with infrared communication systems. Eye safety is a concern because of a combination of two effects. First, the cornea is transparent from the near violet to the near IR. Hence, the retina is sensitive to damage from light sources transmitting in these bands. Secondly, however, the near IR is outside the visible range of light, and so the eye does not protect itself from damage by closing the iris or closing the eyelid. Eye safety can be ensured by restricting the transmit beam strength according to IEC or ANSI standards [7,8].

Skin safety is also a possible concern. Possible shortterm effects such as heating of the skin are accounted for by eye safety regulations (since the eye requires lower power levels than does the skin). Long-term exposure to IR light is not a concern, as the ambient light sources are constantly submitting our bodies to much higher radiation levels than these communication systems do.

3. COMMUNICATIONS DESIGN

Equally important for achieving the design goals of wireless infrared systems are communications issues. In particular, the modulation signal format together with appropriate error control coding is critical to achieving power efficiency. Channel characterization is also important for understanding performance limits.

3.1. Modulation Techniques

To understand modulation in IM/DD systems, we must look again at the channel model $% \left[{{\left[{{{\rm{DD}}} \right]}_{\rm{TO}}} \right]_{\rm{TO}}} \right]$

$$Y(t) = X(t) \otimes c(t) + N(t)$$

and consider its particular characteristics. First, since we are using intensity modulation, the channel input X(t) is optical intensity and we have the constraint $X(t) \ge 0$. The

average transmitted optical power P_T is the time average of X(t). Our goal is to minimize the transmitted power required to attain a certain probability of bit error P_e , also known as a bit error rate (BER).

It is useful to define the signal-to-noise ratio (SNR) as

$$\text{SNR} = \frac{R^2 H^2(0) P_t^2}{R_b N_0}$$

where H(0) is the DC gain of the channel, i.e., it is the Fourier transform of h(t) evaluated at zero frequency, so

$$H(0) = \int_{-\infty}^{\infty} h(t) \, dt.$$

The transmitted signal can be represented as

$$X(t) = \sum_{n=-\infty}^{\infty} s_{a_n} (t - nT_s).$$

The sequence $\{a_n\}$ represents the digital information being transmitted, where a_n is one of L possible data symbols from 0 to L - 1. The function $s_i(t)$ represents one of L pulseshapes with duration T_s , the symbol time. The data rate (or bit rate) R_b , bit time T, symbol rate R_s , and symbol time T_s are related as follows: $R_b = 1/T$, $R_s = 1/T_s$, and $T_s = \log_2(L)T$.

There are three commonly used types of modulation schemes: on/off keying (OOK) with non-return-tozero (NRZ) pulses, OOK with return-to-zero (RZ) pulses of normalized width δ (RZ- δ), and pulse position modulation with L pulses (L-PPM). OOK and RZ- δ are simpler to implement at both the transmitter and receiver than L-PPM. The pulse shapes for these modulation techniques are shown in Fig. 3. Representative examples of the resulting transmitted signal X(t) for a short data sequence are shown in Fig. 4.

We compare modulation schemes in Table 1 by looking at measures of power efficiency and bandwidth efficiency. Bandwidth efficiency is measured by dividing the zerocrossing (ZC) bandwidth by the data rate. Bandwidthefficient schemes have several advantages-the receiver and transmitter electronics are cheaper, and the modulation scheme is less likely to be affected by multipath distortion. Power efficiency is measured by comparing the required transmit power to achieve a target probability of error P_e for different modulation techniques. Both RZ- δ and PPM are more power-efficient than OOK, but at the cost of reduced bandwidth efficiency. However, for a given bandwidth efficiency, PPM is more power-efficient than RZ- δ , and so PPM is most commonly used. OOK is most useful at very high data rates, say 100 Mbps (megabits per second) or greater. Then, the effect of multipath distortion is the most significant effect and bandwidth efficiency becomes of paramount importance [9].

3.2. Error Control Coding

Error control coding is an important technique for improving the quality of any digital communication system. We concentrate here on forward error correction channel



Figure 3. The pulse shapes for OOK, RZ-0.25, and 4-PPM.



Figure 4. The transmitted signal for the sequence 010011 for OOK, RZ-0.25, and 4-PPM.

coding, as this specifically relates to wireless infrared communications; source coding and ARQ (automatic repeat request) coding are not considered here.

Trellis-coded PPM has been found to be an effective scheme for multipath infrared channels [10,11]. The key technique is to recognize that although on a distortionfree channel, all symbols are orthogonal and equidistant in signal space, this is not true on a distorting channel. Hence, trellis coding using set partitioning designed to separate the pulse positions of neighboring symbols is an effective coding method. Coding gains of 5.0 dB electrical have been reported for rate $\frac{2}{3}$ -coded 8-PPM over uncoded 16-PPM, which has the same bandwidth [11].

3.3. Channel Impulse Response Characterization

Impulse response characterization refers to the problem of understanding how the impulse response c(t) in Eq. (1) depends on the location, size, and orientation of the receiver and transmitter. There are basically three classes of techniques for accomplishing this: measurement, simulation, and modeling. Channel measurements have been described in several studies [2,9,12], and these form the fundamental basis for understanding the channel properties. A particular study might generate a collection of hundreds or thousands of example impulse responses $c_i(t)$ for configuration *i*. The collection of measured impulse responses $c_i(t)$ can then be studied by looking at scatterplots of path loss versus distance, scatterplots of delay spread versus distance, the effect of transmitter and receiver orientations, robustness to shadowing, and so on.

Simulation methods have been used to allow direct calculation of a particular impulse response based on a site-specific characterization of the propagation

Table 1. Comparison of Modulation Schemes on Ic	leal Channels
---	---------------

Modulation Type	P_e	ZC Bandwidth
On/off keying (OOK)	$Q(\mathrm{SNR}^{1/2})$	R_b
OOK RZ- δ	$Q(\delta^{-1/2}~{ m SNR}^{1/2})$	$\frac{1}{\delta}R_b$
L-PPM	$Q\left((0.5L * \log_2(L))^{1/2} \mathrm{SNR}^{1/2} ight)$	$rac{L}{\log_2 L} R_b$

2929

environment [13,14]. The transmitter, the receiver, and the reflecting surfaces are described and used to generate an impulse response. The basic assumption is that most interior surfaces reflect light diffusely in a Lambertian pattern, that is, all incident light, regardless of incident angle, is reflected in all directions with an intensity proportional to the cosine of the angle of the reflection with the surface normal. The difficulty with existing methods is that accurate modeling requires extensive computation.

A third technique attempts to extract knowledge gained from experimental and simulation-based channel estimations into a simple-to-use model. In Ref. 15, for example, a model using two parameters (one for path loss, one for delay spread) is used to provide a general characterization of all diffuse IR channels. Methods for relating the parameters of the model to particular room characteristics are given, so that system designers can quickly estimate the channel characteristics in a wide range of situations.

4. STANDARDS AND SYSTEMS

We examine the details of the two dominant wireless infrared technologies, IRDA and IEEE 802.11, and other commercial applications.

4.1. Infrared Data Association Standards (IRDA)

The Infrared Data Association [16], an association of about 100 member companies, has standardized low-cost optical data links. The IRDA link transceivers or "ports," appear on many portable devices, including notebook computers, personal digital assistants, and also computer peripherals such as printers.

The series of IRDA transmission standards are described in Table 2. The current version of the physicallayer standards is IrPHY 1.3. Data rates ranging from 2.4 kbps to 4 Mbps are supported. The link speed is negotiated by starting at 9.6 kbps.

Most of the transmission standards are for short-range, directed links with an operating range from 0 to 1 m. The transmitter half-angle must be between 15° and 30° , and the receiver field-of-view half-angle must be at least 15° . The transmitter must have a peak-power wavelength between 850 and 900 nm.

4.2. IEEE 802.11 and Wireless LANs

The IEEE has published a set of standards for wireless LANs, IEEE 802.11 [17]. The IEEE 802.11 standard is designed to fit into the structure of the suite of IEEE 802 LAN standards. Hence, it determines the physical layer (PHY) and medium-access control layer (MAC) leaving the logical link control (LLC) IEEE to 802.2. The MAC layer uses a form of carrier-sense multiple access with collision avoidance (CSMA/CA).

The original standard supports both radio and optical physical layers with a maximum data rate of 2 Mbps. The IEEE 802.11b standard adds a 2.4-GHz radio physical layer at up to 11 Mbps and the IEEE 802.11a standard adds a 5.4-GHz radio physical layer at up to 54 Mbps.

The two supported data rates for infrared IEEE 802.11 LANs are 1 and 2 Mbps. Both systems use PPM but share a common chip rate of 4 Mchips/s, as explained below. Each frame begins with a preamble encoded using 4 Mbps OOK. In the preamble, a 3-bit field indicates the transmission type, either 1 or 2 Mbps (the six other types are reserved for future use). The data are then transmitted at 1 Mbps using 16-PPM or 2 Mbps using 4-PPM. 16-PPM carries $\log_2(16)/16 = \frac{1}{4}$ bits/chip, and 4-PPM carries $\log_2(4)/4 = \frac{1}{2}$ bit/chip, resulting in the same chip time for both types.

The transmitter must have a peak-power wavelength between 850 and 950 nm. The required transmitter and receiver characteristics are intended to allow for reliable operation at link lengths up to 10 m.

4.3. Building-to-Building Systems

Long-range (>10 m) infrared links must be directed LoS systems in order to ensure a reasonable path loss. The emerging products for long-range links are typically designed to be placed on rooftops [18,19], as this provides the best chance for establishing line-of-sight paths from one location to another in an urban environment. These high-data-rate connections can then be used for enterprise network access or metropolitan- or campus-area networks.

Several design issues are specific to these systems that are unique to these long-range systems [3]: (1) *atmospheric path loss*, which is a combination of clean-air absorption from the air and absorption and scattering from particles in the air, such as rain, fog, and pollutants;

Table 2. Http://bata Transmission Standards							
Version	Link Type	Link Range (m)	Data Rate	Modulation			
1.3	Point-to-point	1	2.4–115.2 kbps	$RZ - \frac{3}{16}$			
1.3	Point-to-point	1	$576 \mathrm{kbps}$	$RZ-\frac{1}{4}$			
			$1152 \rm \ kbps$	$RZ-\frac{1}{4}$			
1.3	Point-to-point	1	4 Mbps	4-PPM			
$VFIR^{a}/1.4$	Point-to-point	1	16 Mbps	OOK			
$AIR^b/proposed$	Network	4	4 Mbps	_			
		8	$250 \mathrm{kbps}$	_			

 a VIFR = Very Fast Infrared.

 b AIR = Advanced Infrared.

(2) *scintillation*, which is caused by temperature variations along the LoS path, causing rapid fluctuations in the channel quality; and (3) *building sway*, which can affect alignment and result in signal loss unless the transceivers are mechanically isolated or active alignment compensation is used.

4.4. Other Applications

Wireless infrared communication has found several markets in and around the home, car, and office that fall outside the traditional telecommunications markets of voice and data networking. These can be classified as either wireless input devices or wireless control devices, depending on one's perspective. Examples include wireless computer mice and keyboards, remote controls for entertainment equipment, wireless videogame controllers, and wireless door keys for home or vehicle access. All such devices use infrared communication systems because of the attractive combination of low cost, reliability, and light weight in a transmitter/receiver pair that achieves the required range, data rate, and data integrity required.

5. TECHNOLOGY OUTLOOK

In this section, we discuss how competition from radio and developments in research will impact the future uses of wireless infrared communication systems.

5.1. Comparison to Radio

Wireless infrared communication systems enjoy significant advantages over radio systems in certain environments. First, there is an abundance of unregulated optical spectrum available. This advantage is shrinking somewhat as the spectrum available for licensed and unlicensed radio systems increases with the modernization of spectrum allocation policies.

Radio systems must make great efforts to overcome or avoid the effects of multipath fading, typically through the use of diversity. Infrared systems do not suffer from time-varying fades because of the inherent diversity in the receiver, thus simplifying design and increasing operational reliability.

Infrared systems provide a natural resistance to eavesdropping, as the signals are confined within the walls of the room. This also reduces the potential for neighboring wireless communication systems to interfere with each other, which is a significant issue for radio-based communication systems.

In-band interference is a significant problem for both types of systems. A variety of electronic and electrical equipment radiates in transmission bands of current radio systems; microwave ovens are a good example. For infrared systems, ambient light, either humanmade (synthetic) or natural, is a dominant source of noise.

The primary limiting factor of infrared systems is their limited range, particularly when no good optical path can be made available. For example, wireless communication between conventional rooms with opaque walls and doors cannot be accomplished; one must resort to using either a radio-based or a wireline network to bypass the obstruction.

5.2. Research Challenges

Various techniques have been considered to improve on the performance of wireless infrared communication systems.

At the transmitter, the radiation pattern can be optimized to improve performance characteristics such as range. Some optical techniques for achieving this are diffusing screens, multiple-beam transmitters, and computer-generated holographic images.

At the receiver, performance is ultimately determined by signal collection (limited by the size of the photodetector) and by ambient noise filtering. Optical interference filters can be used to reduce the impact of background noise; the primary difficulty is in achieving a wide field of view. This can be done using nonplanar filters or multiple narrow-FoV receiving elements.

Some recent developments and research programs are described in Ref. 20, and an on-line resource guide is maintained in Ref. 21.

6. CONCLUSIONS

Wireless infrared communication systems provide a useful complement to radio-based systems, particularly for systems requiring low cost, lightweight, moderate data rates, and only requiring short ranges. When LoS paths can be ensured, range can be dramatically improved to provide longer links.

Short-range wireless networks are poised for tremendous market growth in the near future, and wireless infrared communications systems will compete in a number of arenas. Infrared systems have already proved their effectiveness for short-range temporary communications and in high-data-rate, longer-range point-to-point systems. It remains an open question whether infrared will successfully compete in the market for general-purpose indoor wireless access.

BIOGRAPHY

Jeffrey B. Carruthers received his B.Eng. degree in computer systems engineering from Carleton University, Ottawa, Canada, in 1990, and M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, Berkeley, California, in 1993 and 1997, respectively. Since 1997, he has been an assistant professor in the Department of Electrical and Computer Engineering at Boston University, Boston, Massachusetts. Previously, he was with SONET Development Group of Bell-Northern Research, Ottawa, Canada, from 1990 to 1991. From 1992 to 1997 he was a research assistant at the University of California, Berkeley. Dr. Carruthers was an NSERC 1967 Postgraduate Scholar, and he received the National Science Foundation CAREER Award in 1999. His research interests include wireless infrared communications, wireless networking, and digital

communications. He is a member of the IEEE and the IEEE Communications Society.

BIBLIOGRAPHY

- F. R. Gfeller and U. H. Bapst, Wireless in-house data communication via diffuse infrared radiation, *Proc. IEEE* 67: 1474–1486 (Nov. 1979).
- J. M. Kahn and J. R. Barry, Wireless infrared communications, *Proc. IEEE* 85: 265–298 (Feb. 1997).
- 3. D. Heatley, D. Wisely, I. Neild, and P. Cochrane, Optical wireless: The story so far, *IEEE Commun. Mag.* 72–82 (Dec. 1998).
- 4. J. R. Barry, Wireless Infrared Communications, Kluwer, Boston, 1994.
- J. R. Barry and J. M. Kahn, Link design for non-directed wireless infrared communications, *Appl. Opt.* 34: 3764–3776 (July 1995).
- R. Narasimhan, M. D. Audeh, and J. M. Kahn, Effect of electronic-ballast fluorescent lighting on wireless infrared links, *IEE Proc.-Optoelectron.* 143: 347–354 (Dec. 1996).
- 7. International Electrotechnical Commission, CEI/IEC 825-1: Safety of Laser Products, 1993.
- ANSI-Z136-1, American National Standard for the Safe Use of Lasers, 1993.
- 9. J. M. Kahn, W. J. Krause, and J. B. Carruthers, Experimental characterization of non-directed indoor infrared channels, *IEEE Trans. Commun.* **43**: 1613–1623 (Feb.-April 1995).
- D. Lee and J. Kahn, Coding and equalization for PPM on wireless infrared channels, *IEEE Trans. Commun.* 255-260 (Feb. 1999).
- D. Lee, J. Kahn, and M. Audeh, Trellis-coded pulse-position modulation for indoor wireless infrared communications, *IEEE Trans. Commun.* 1080–1087 (Sept. 1997).
- H. Hashemi et al., Indoor propagation measurements at infrared frequencies for wireless local area networks applications, *IEEE Trans. Vehic. Technol.* 43: 562-576 (Aug. 1994).
- J. R. Barry et al., Simulation of multipath impulse response for indoor wireless optical channels, *IEEE J. Select. Areas Commun.* 11: 367-379 (April 1993).
- M. Abtahi and H. Hashemi, Simulation of indoor propagation channel at infrared frequencies in furnished office environments, *IEEE International Symposium on Personal*, *Indoor, and Mobile Radio Communications (PIMRC)* 1995, 306-310.
- J. B. Carruthers and J. M. Kahn, Modeling of nondirected wireless infrared channels, *IEEE Trans. Commun.* 1260-1268 (Oct. 1997).
- 16. http://www.irda.gov.
- IEEE Standard 802.11-1999, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, 1999.
- 18. http://www.canobeam.com.
- 19. http://www.terabeam.com.
- 20. Special issue of *IEEE Communications Magazine* on Optical Wireless Systems and Networks (Dec. 1998).
- 21. http://www.bu.edu/wireless/irguide.

WIRELESS IP TELEPHONY

DAVID R. FAMOLARI Telecordia Technologies Morristown, New Jersey

1. INTRODUCTION

No two innovations have done more in the 1990s to advance communications than the wireless cellular telephone network and the Internet. Wireless cellular telephony extended access to the Public Switched Telephone Network (PSTN), the conventional landline telephone system, to users with small handheld radio terminals. These users could then establish and maintain voice calls anywhere within radio coverage, which for practical purposes in the United States is nationwide. This ushered in an era of mobile communications that disassociated a telephone user with a geographic location. No longer did dialing a particular number ensure that a caller would find (or not find) the intended recipient "at home," "at work," or elsewhere; with wireless telephony they could in fact be anywhere. And so could the caller. The desire to be mobile, coupled with falling subscription charges and rising voice quality, motivate an increasing number of customers to flock to the cellular telephone network. The reliability and quality of wireless phone service has even prompted some to forego landline service all together. Simply put, wireless telephony has made good on the promise of anywhere/anytime voice communications.

Based on the unifying, packet switching Internet Protocol (IP) [1], the Internet has had no less a dramatic impact on how society communicates. Its advent has interconnected devices over vast distances at limited costs and made possible rich multimedia content exchanges, most notably with the emergence of the World Wide Web (WWW). The open and standard protocols of the Internet allow equipment vendors to easily produce infrastructure products that interoperate with other vendor's products, increasing competition and fostering economies of scale. This has led to the pronounced deployment of Internet architecture throughout the world and, consequently, extended the reach of the Internet on a global scale. Because of its reach and its ability to support a variety of services, the Internet has evolved from a loose interconnection of computers employed by researchers to exchange files into a global network infrastructure that is an essential underpinning to commercial, governmental, and private communication.

Both the Internet and wireless telephony have created unprecedented operating freedoms and are forcing paradigm shifts in business and social communication practices. Their impact, however, has emerged from opposite design philosophies. The wireless telephone network provides ubiquitous access to speech applications; the Internet, on the other hand, delivers a wide variety of application types to fixed locations. The freedom and mobility offered by wireless networks provides a perfect compliment to the economies of scale and flexibility offered by the IP-based Internet.

2932 WIRELESS IP TELEPHONY

Now, at the beginning of the twenty-first century, the two most defining communications breakthroughs of the late twentieth century are starting to merge. As IP becomes an important driver of both core and access networks, it must support wireless voice at comparable levels of spectrum efficiency and voice quality as cellular telephony. Similarly, wireless network practices must be modified to accommodate IP-based packet protocols. It is our aim here to describe the design challenges and technical hurdles facing successful deployment of wireless IP telephony.

2. BACKGROUND AND BASIC CONCEPTS

The design goals of the current breed of wireless telephony standards, commonly referred to as *second*-generation $(2G)^1$ standards, were to carry digital voice conversations to and from the PSTN with packet data networking only as an afterthought. As a consequence, current cellular systems have only limited packet data capabilities, usually to exchange brief text messages such as the Short Message Service (SMS). However new breeds of cellular communications systems are promising advanced features for packet data networking, along with faster transmission speeds and higher capacity. These new communications systems are widely referred to as *third-generation* (3G) systems. It is in 3G systems that the first steps are being taken towards integrating IP transport with wide-area wireless networks.

The new breed of 3G systems were initiated by the International Telecommunications Union (ITU), a United Nations governing body, under a directive called "International Mobile Telecommunications for the year 2000" (IMT-2000). IMT-2000 is a family of guideline specifications and recommendations serving as an umbrella standard to harmonize the evolution of the disparate 2G radio technologies currently in place; most notably the CDMA-based IS95 systems in the United States and the GSM systems in Europe and elsewhere. Regionalized activities within the IMT-2000 family emerged based on these two technologies. These include the Third Generation Partnership Project (3GPP), centered on evolving the GSM standard under the name Universal Mobile Telecommunications Service (UMTS) [2], and the Third Generation Partnership Project 2 (3GPP2), focused on evolving the IS95 standard, referred to as "cdma2000" [3].² Both groups are actively mapping out architectures and protocol references to support high-rate packet data services that will be used to carry IP telephony over their respective networks.

The challenges that these groups will face are to accommodate three fundamental issues in successful

delivery of wireless voice. These issues are service quality of the perceived voice output stream, efficient use of the band-limited medium, and mobility of wireless users.

1. Service Quality. Wireless telephone systems have been efficiently designed to carry one type of application: voice, a time-sensitive, error-tolerant communication service. Its perceived quality degrades more rapidly when the delay characteristics of the underlying delivery change than as the error characteristics change. The destination of each call is usually a human being who can compensate for imperfections in the received signal. This is akin to deriving the proper meaning of a sentence when a word or two is mispronounced. When sounds arrive with noticeable variation in their inter-arrival times (called *jitter*), or when there are long lapses in the conversation (delay), conversation can become tedious. Circuit switched principles, at the core of wireless telephone systems, are well suited for voice since they offer timely and regular access to the transport medium. The Internet, on the other hand, thrives on the principles of packet switching. Consequently IP packets often take different routes through the network and can arrive out of sequence with variable delays. Therefore the basic IP protocols are not well suited to deliver time-sensitive applications such as voice.

2. Efficiency. Wireless telephony contends with limited spectrum and an unpredictable, time-varying physical channel that is subject to path loss, fading, and multipleaccess interference. In order to support high capacity and revenue, service providers must make best utilization of their given spectrum. To this end, cellular telephony employs techniques that make best use of scarce, unreliable resources to deliver near-toll-quality voice. These techniques include employing low-bit-rate voice codecs to compress speech and reducing frame sizes to make voice packets less vulnerable to rapid fluctuations in the channel. The protocols and architecture of the Internet give prominence to end-to-end principles over centralized approaches. While offering flexibility, this demands that each packet contain a greater degree of control information than circuit-switched packets. The additional control information is embedded in the detailed protocol headers that are included in each IP data packet. Consequently, squeezing these heavyweight protocol headers into the relatively small packet sizes of the wireless channel gives rise to a number of performance problems related to efficiency.

3. *Mobility*. A fundamental design concept of wireless telephony is to allow users to seamlessly change their point of attachment to the network at any time. This must occur without explicit reconfiguration or significant performance loss. IP routing, however, associates an IP address with a fixed attachment to a router. When this association is no longer valid, as is the case when wireless users move, standard IP routing procedures are unable to deliver packets to and from the mobile terminal. Furthermore, re-establishing a valid association under the standard IP procedure may require manual intervention and explicit reconfiguration that will terminate any ongoing communications. Therefore these routing protocols require

¹ Second-generation systems represented a significant leap from the first generation Analog Mobile Phone System (AMPS), including the use of digital modulation techniques, enhanced security features, better spectral efficiency and longer mobile terminal battery life.

 $^{^2}$ For more detail on 3G wireless standards, the reader is referred to Ref. 4.

extensions and additional control architectures to transfer associations in midsession. In other words, mobility solutions for IP will need to offer seamless transfers in order to support wireless voice effectively.

In the remainder of this chapter we focus on each of these three principles and discuss modifications to the IP protocols and architecture to successfully offer IP telephony services.

3. SERVICE QUALITY FOR WIRELESS IP TELEPHONY

Service quality can be characterized by three parameters; packet loss, delay, and delay variability. In the wired network, heavy traffic at ingress points can overload IP routers, causing packets to be dropped. Even in the absence of heavy traffic, the detrimental effects of the wireless channel can corrupt voice packets so that they cannot be recovered at the receiver and are considered lost. Packet loss is therefore a considerable problem on wireless links. One-way delay is the time from when a voice packet enters the encoder at the source terminal to when it exits the decoder at the destination terminal. As mentioned earlier, the variability in packet arrival times is referred to as jitter. These three parameters are interdependent and all contribute to the overall service quality. In wireless networks where packet loss can be high, delay and jitter will increase correspondingly. Human perceptions can accommodate reasonable delays, but service quality becomes degraded when those delays have a high variation from packet to packet.

Wireless IP telephony service quality is dependent upon two issues: the service quality that voice receives over the air interface and the service quality that it receives while traversing the wired backbone network of the service provider. *Wired IP service quality* management has been a topic of much research. Many efforts have been geared toward supporting classes of service above and beyond the best-effort service of the typical Internet. These technologies attempt to provide more reliable delay and jitter performance by classifying traffic and giving priority treatment to certain traffic classes. We focus here on the service quality aspects particular to *wireless* transmission and refer the reader to Refs. 5–7 for more detailed information on wired service quality measures.

3.1. Wireless Service Quality

3.1.1. Link-Layer Solutions. Information transfer over wireless links is subject to impairments and environmental constraints that landline communications are free from. Factors affecting the wireless channel include interference, shadowing, path loss, and fading. These affects can have dramatic negative impacts on the signal-to-interference ratio (SIR). Lower SIR levels lead to increases in the frame error rates (FERs). Higher FER, in turn, increases packet loss. Furthermore, these wireless channel effects vary over small distances making signal quality very sensitive to terminal mobility. Simply increasing the bit rate cannot overcome the impediments of the wireless channel. Though all 3G systems will offer much higher

available bandwidths for packet data [supporting rates of 144 kbps (kilobits per second) at high mobility rates, 384 kbps at pedestrian mobility rates, and 2 Mbps for indoor systems], voice service quality is still primarily dependent on achieving timely delivery of good-quality voice packets.

In order to improve the quality of the voice packets, wireless networks have focused on making voice packets more robust to transmission errors. In addition to physicallayer solutions that seek to improve the SIR performance of receivers, wireless system providers employ coding techniques such as forward error correction (FEC). FEC allows the receiver to locally reconstruct packets corrupted by bit errors without forcing retransmissions [8]. This scheme works by chopping frames into a smaller number of codewords and inserting a series of parity bits into each codeword that helps the receiver recover from a small number of bit errors; the greater the number of parity bits, the greater the recovery ability. Inherent in FEC, therefore, is a tradeoff between efficiency and resiliency. A resilient codeword contains more parity bits than does a less resilient one and thus has more transmission overhead, that is, less efficient use of the limited bandwidth. On the other hand, a less resilient code has a higher probability of unrecoverable error and can lead to retransmissions. This is detrimental in both efficient use of bandwidth and delay performance. Typical coding rates—the ratios of information bits to the total size of the codeword — used in wireless systems are usually on the order of $\frac{1}{4}$ to $\frac{2}{3}$. Some systems even offer dynamic FEC coding strategies that change the coding rate on the basis of the perceived error characteristics of the channel. The proper use of FEC techniques, therefore, can improve the integrity of the transmission sequence and reduce delays incurred on error-prone wireless links by reducing retransmissions.

The effectiveness of FEC coding is improved by interleaving [9]. Often wireless channels do not exhibit statistically independent error properties from one bit to the next. This means that bit errors tend to be lumped together in bursts. FEC codes can correct only a small number of errors per codeword. As such, they can be ineffective when dealing with very bursty error channels. Interleaving works to scramble the bit orderings before transmission and then place them back in proper order at the receiver before decoding. Thus bits that travel consecutively over the air will not be consecutive when presented to the FEC decoder. When deep fades corrupt a series of bits in transit, the interleaving process at the receiver will disperse those errors throughout the frame over multiple codewords, improving the effectiveness of the FEC scheme. This again implies that more packets can be corrected without retransmissions thereby reducing delay.

Another FEC technique employed in wireless voice systems to ensure good-quality characteristics is unequal error protection [10]. Unequal error protection is the practice of classifying certain bits in the voice frame as more important then other bits. If such essential bits are corrupted in transmission, then the frame cannot be used because too much fundamental information about the voice sample will have been lost. These bits, as a result, are protected with FEC codes while the others are not. The nonessential bits in the voice sample can be delivered to the higher layers even if they contain bit errors. Although this reduces the sound quality of the voice sample, the end effect is still tolerable for the listener. The benefit of this approach is twofold: (1) the overhead associated with FEC coding is reduced and (2) usable (albeit less-than-perfect) voice information is delivered in cases where it would otherwise have been dropped or retransmitted.

In order to combat jitter and ensure smooth playback, real-time media systems buffer received voice packets. This solution to combat poor delay performance exists in both the wireless and wired telephone systems. With a buffered voicestream the receiver presents packets to the decoder at regular intervals even if they arrive at irregular intervals. Buffering at the receiver also helps establish correct packet ordering by allowing the receiver to collect a number of unordered packets and place them in the proper sequential order before they are played. This concept is illustrated in Fig. 1. Jitter buffers, however, add to the one-way delay already incurred due to packet encoding and transmission. Jitter buffer lengths must be designed with this delay penalty in mind. The International Telecommunications Union (ITU) states that one-way delay for voice samples must be lower than 400 ms for acceptable voice quality for almost all applications (with the exception of certain long-haul satellite communications). Furthermore, the ITU recommends that delays be below 150 ms [11]. Jitter buffer lengths are an important piece of the overall delay budget that includes codec, medium access, network, and transmission delays.

In addition to these techniques, all 3G systems will also have advanced methods for guaranteeing dataflows greater degrees of access to their air interfaces. Most notably these include forms of traffic classification and priority scheduling at the MAC layer that can dedicate resources to real-time packet streams [12]. These guarantees often provide a minimum bit rate and/or delay. Additionally, they provide intelligent voice-friendly queue management policies. The UMTS, in particular, has specified two real-time traffic classes, conversational and streaming [13]. The conversational class is the most delay-sensitive and is expected to handle wireless IP telephony. Certain deliverability attributes are associated with each class, such as maximum and guaranteed bit rates, maximum transfer delays, handling priorities, and whether packets with errors should be forwarded to higher layers. Using these attributes to define the service quality requirements for various traffic flows will allow 3G



Figure 1. Jitter buffering.

wireless systems to service time-sensitive voice packets before delay-tolerant data packets.

3.1.2. Transport-Layer Solutions. While these link-level improvements will go a long way toward improving the service quality of wireless voice, performance is also highly dependent on the transport protocols used to deliver samples from the mobile terminal to their final destination. Reliable end-to-end transport protocols that promise ordered delivery of error-free packets, such as Transmission Control Protocol (TCP) [14], respond to link-level errors by requesting retransmissions. Such retransmissions hold up the timely delivery of subsequent packets. Furthermore, TCP operates on a principle that packet loss is due to congestion in the network and will thus respond to errors on the channel by generating less traffic. TCP will at first throttle the bit rate and then slowly increase it as the signs of congestion dissipate. This has dire consequences on the performance of voice applications by slowing down the service unnecessarily and imposing unacceptable delays. These issues are further exacerbated in wireless networks where link-level errors are frequent and seldom the result of congestion. It is clear that the application and performance characteristics of wireless voice are at odds with the design goals of TCP. The emphasis on timely delivery of voice over error-free reception cannot be reconciled with the TCP design philosophy, which sacrifices delays to achieve an error-free result.

As a result, the use of TCP to carry wireless voicestreams is not likely. User Datagram Protocol (UDP) [15], the connection-less alternative to TCP, in conjunction with the Real-Time Transport Protocol (RTP) [16], will be the most prominent implementation of voice over IP. UDP requires no retransmissions and is not session based, meaning that it carries no state and treats each packet individually independent of packets before or after it. This statelessness is not necessarily well suited for isochronous media where many packets are often carried with similar characteristics in a stream and it is advantageous to make use of state information. However, protocols such as RTP, which are used in conjunction with UDP, provide functionality to real-time voice above what simple UDP can provide.

RTP was developed to use the basic UDP datagram in order to provide end-to-end support for time-sensitive applications. RTP offers no congestion control and no promises of reliability; however, in the world of voice this is a boon. Without reliability and congestion control, the transport protocol will not mistake link-level errors for congestion and will not delay packets while waiting for retransmissions. End devices can use the information provided in RTP packets to determine the real-time characteristics of the received packet. RTP provides a timestamp field that allows receivers to determine whether an incoming packet is "fresh" and should be played or is "stale" and should be dropped. Sequence numbers are used to identify gaps in the reception of packets. This may signal that an error has occurred or that packets have been received out of order. RTP provides the necessary information to the receiver to

reconstruct the original stream when packets are received out of sequence. RTP is also particularly well suited for delivering multimedia traffic and provides mechanisms by which media streams can be synchronized and multiplexed, such as audio and video for a videoconference. RTP is by far the most frequently used protocol in the wired Internet for transferring real-time information. Because of its widespread use and special features designed for real-time traffic, RTP is the natural transport protocol for wireless IP telephony.

Wireless IP telephony providers must strive to offer service qualities that are comparable to the cellular voice services that customers are accustomed to. Strict requirements on delay, jitter, and packet loss create a challenge in ensuring service quality. In addition to intelligent service quality management of wired backbone links, this challenge will have to be met by coordinated efforts between radio-level mechanisms operating over the wireless link and transport protocols operating end to end.

4. EFFICIENCY AND WIRELESS IP TELEPHONY

Wide-area wireless channels typically employ bandwidths much smaller than those found on landline networks. Furthermore the unpredictable nature of the wireless channel makes the use of small link-level packet sizes advantageous. Small packet sizes are less vulnerable to channel fluctuations and help the radio network recover more gracefully from packet losses. As an example, typical cellular systems today, as well as future 3G systems, employ basic packet sizes that are on the order of 20 ms.

These small packet sizes represent a major challenge to the use of IP-type protocols that employ large headers. The resulting overhead can have detrimental effects on the efficiency of the system. Efficient use of wireless resources allows service providers to support higher capacities. For wireless IP telephony providers to match the level of spectral efficiency of traditional cellular networks, the resultant overhead of IP transport must be reduced. Header compression techniques are the most effective way to reduce IP packet overheads, and several such compression schemes have been defined for compressing a variety of protocol types. Below we discuss the most relevant aspects of header compression to wireless IP telephony.

4.1. Header Compression

As indicated earlier, voice over IP networks will be supported by the Real-Time Transport Protocol, which runs over UDP. Thus the typical protocol layering for IP voice looks like RTP/UDP/IP.³ Each of these protocols introduces its own overheads in the form of required headers. Typically this value totals 40 bytes, including 20 bytes for RTP, 8 bytes for UDP, and 12 bytes for IP. The protocol header fields for RTP/UDP/IP are shown in Fig. 2, where the numbers represent bit positions and are used to indicate the length of the header fields. We discuss only a few of the header fields below; for more detailed information on the RTP/UDP/IP header fields the reader is referred to the literature [1,15,16].

This 40-byte RTP/UDP/IP overhead represents a significant portion of available wireless bandwidth. Current wireless voice packetization schemes, including

³ This is taken to mean that RTP is on top of UDP, which is on top of IP, indicating that a packet must travel down the protocol stack, traversing first the RTP layer, then the UDP and IP layers. Other references write the protocol layering as IP/UDP/RTP, indicating that IP is outside UDP that is outside RTP. This latter notion emphasizes that IP protocol headers are followed by UDP and RTP protocol headers.



Figure 2. RTP/UDP/IP protocol headers.

the ITU standard G.729 [17], employ 8-kbps codecs where 20 bytes of voice information is sent every 20 ms. Therefore voice packets carried under the RTP/UDP/IP regime will incur a 200% overhead price. Clearly deep RTP/UDP/IP header compression is required for wireless IP telephony to achieve the efficiency that cellular operators need to service a growing number of subscribers.

4.1.1. Compression State. Header compression works on the principle that many RTP/UDP/IP packet header fields change either predictably or very infrequently from packet to packet. For example, the IP addresses of the two correspondents never change during the course of a typical session; however, the 64 bits (8 bytes) used to denote source and destination IP addresses, is included in each UDP header. A similar situation exists for the source and destination port addresses, which account for 2 bytes. Addressing and port assignments, which remain largely static over the lifetime of a call, account for roughly 25% of the total overhead. Header compression schemes leverage the predictability of packet headers from packet to packet to achieve compression levels on the order of 95-97%; in some instances they reduce 40-byte overhead to 1 or 2 bytes.

Header compression schemes are able to achieve these levels of compression by first eliminating wellknown a priori information or information that can be inferred from other mechanisms, such as the link layer. Fields such as the IP and RTP version numbers are expected to be well known, other values such as the IP header and payload lengths can be successfully inferred from the link layer. Moreover, sending non-a priori, but static, information only once at the onset of the session reduces overhead considerably. Values such as the abovementioned 8-byte IP addresses and 2-byte port numbers, among others, can be sent once and will remain constant over a large number of packet headers. Sending these values once helps the compressor and decompressor establish a *context*, or compression state, by which future compressed packets can be evaluated. Compression state is the knowledge necessary to reconstitute the full header from a compressed header. After the initial sending, it is necessary to only provide delta values, or updates, in header fields that change instead of the absolute values. This significantly reduces the amount of overhead transmission.

Figure 3 shows the general architecture of a header compression scheme where a *compressor* takes full RTP/UDP/IP headers and generates compressed headers that are sent over the wireless channel. On the receiving side the decompressor attempts to reconstruct the original header by applying the corresponding decompression scheme using the current header context. When the context becomes lost or loses synchronization between the compressor and the decompressor, as is the result of link-layer errors, the compression scheme must be able to restore context.

4.1.2. Error Recovery. An important consideration in the design of a header compression scheme for the wireless environment is its ability to recover from errors. Cassner



Figure 3. Header compression architecture.

and Jacobson proposed the earliest header compression technique for RTP/UDP/IP, compressed RTP (CRTP), [18] which became a draft standard in 1999. While this scheme worked well for telephone dialup connections and other low-loss link layers, its design did not perform well in a highly variable and error-prone wireless channel. When packets were received in error, the header compression states at the compressor and the receiver would lose context and full packet headers had to be transferred to reestablish synchronized header state. The effect of long round-trip times, as is the case with wireless IP telephony, has dramatic effects on the error recovery performance. When links have long round-trip times the context cannot be regained as quickly and many compressed packets will require retransmissions or be dropped. Performance lags in CRTP over wireless links became evident [19], and efforts were made to design more robust header compression schemes that could adapt to imperfections in the channel and gracefully recover from errors.

4.1.3. Robust Header Compression. To meet the error performance requirements necessitated by volatile wireless links, header compression schemes have to be reliable and robust. The *robust header compression* (ROHC) scheme [20] was created to make header context less sensitive to packet loss and delay as well as make context recovery faster. This approach repairs context locally, thereby eliminating the need to send update information over the wireless link.

The key factor in ROHC is that cyclic redundancy check (CRC) codes are computed on the uncompressed headers and are sent along with the compressed header. CRC codes are the result of passing a string of bits through a generator polynomial. The CRC codes are then sent along with the bits used to generate the codes. The receiver applies the same function to the received bit string, generating a local copy of the CRC code. If the local copy and the received CRC code are not equal, the receiver is sure that a transmission

error has occurred. CRC codes provide a reliable errordetection mechanism. The decompressor, after generating its copy of the full header from the received compressed header, will perform a CRC check on the full header. This allows the decompressor to reliably determine whether the decompression process was successful. In addition to being able to repair context locally, ROHC is also capable of withstanding a number of consecutive packet errors without losing context. This is important as wireless environments seldom have bit-independent channels and errors often arrive in bursts. ROHC can support upwards of 24 consecutive packet errors without losing context [21].

4.1.4. Performance of Header Compression. Header compression schemes can be evaluated along three distinct performance attributes: compression efficiency, robustness, and compression reliability. In terms of compression efficiency, ROHC can optimally reduce the 40-byte RTP/UDP/IP header to a single byte. Under bit error rates typical of wireless channels, it can achieve an average of 2.27 bytes of overhead [22]. Furthermore, compression efficiency is enhanced by the ability to restore context locally, reducing the transmission of noncompressed headers over the wireless link. Compression reliability, the ability to ensure that decompressed headers are accurate representations of the uncompressed headers, is achieved through the use of CRC codes. This provides a highly reliable way for the decompressor to determine whether the decompressed packet is correct. Finally, ROHC provides a high degree of robustness by correcting loss of context locally and maintaining context in the presence of multiple consecutive errors. Performance results [22] show that under a simulated WCDMA channel operating at a BER of 0.0002 the FER achieved with CRTP is 1.10% while the FER achieved with ROHC was 0.12%. At a higher BER value of 0.001, the frame error rates were 4.06% and 0.81% for CRTP and ROHC, respectively.

It is clear that this new class of robust header compression will be critical to improving the efficiency and performance of wireless IP telephony.

5. MOBILITY

Mobility creates problems with IP routing protocols and can break ongoing sessions. However, wireless IP telephony must be as seamless as present cellular telephony. This requires the ability to change points of attachment to the wireless network while maintaining connectivity with minimal disruption. Cellular solutions address mobility by monitoring channel assignments, code allocations, and received power levels. The introduction of IP transport, however, requires additional mobility solutions that are not addressed by the traditional linklayer cellular mobility techniques. These requirements for *network* mobility extend beyond the *physical*, or *link-level*, mobility offered in cellular networks. The basic functions needed to support mobile access to IP-based networks include

• *Detection* — terminals learn when they have entered new network areas.

- *Registration*—users indicate their presence and requirements to the network.
- *Configuration* network adapts nodes to the particular network characteristics, including IP address assignment and configuration of the default router.
- Authentication, authorization, and accounting (AAA)—validate users and their permission and record their usage for billing and management purposes.
- *Dynamic address binding*—provides a dynamic mapping of old network addresses with new network addresses.

To become a full network participant a user must first have means of physically detecting and connecting to a network. This entails establishing a valid link with the appropriate physical layer protocols, after which a terminal can format information in a contextually meaningful manner. When basic connectivity has been established the mobile and the network can begin to perform parameter negotiations. This occurs during the registration process where the network learns of a terminal's presence and requirements.

Configuration involves fulfilling any registration requests and providing information to enable the mobile to properly orientate itself to the new network surroundings. This may include assigning a new IP address and passing the locations of default routers and network servers. After configuration the next step is for the network to grant the user access to network resources based on AAA measures. The network arrives at these decisions based on negotiations using credentials passed between the terminal and/or user, either explicitly or implicitly. Additionally the network may validate these credentials with third parties located in outside networks. Finally, once the user and terminal have been properly authenticated, dynamic address binding creates an association between the new configuration and the old configuration. This allows mobile terminals to be found after they change networks and allows active sessions to be maintained across different points of attachment transparently.

5.1. IP Mobility for Wireless Telephony

Current cellular networks have mechanisms in place to successfully support link-level mobility. Additionally roaming agreements between service providers allow mobile subscribers from one provider to access another provider's network. In this sense, current cellular networks already have in place mechanisms to support registration, configuration, and AAA functions associated with mobility. However, since mobile telephone numbers are constant and do not change, there is no need to perform dynamic address binding. Also since addressing is valid over the entire network, there is no need to detect when an address change is required. Therefore detecting when address changes are required and dynamically binding those addresses represents a major challenge for current cellular networks to provide mobility.

Supporting dynamic address binding for wireless IP telephony users poses some unique challenges. The IP protocol inherently links physical location with network representation. In other words, an IP address represents a host's physical location on the network as well as its identity within the network. IP uses this association to route packets efficiently to destination addresses. When this association is broken, or invalid, packets can no longer reach their destinations.

5.2. Mobile IP

The industry standard regarding IP mobility is called mobile IP [23] and is actively being designed into the all-IP architectures of next-generation networks including the 3GPP and the 3GPP2 [24]. Mobile IP creates a level of indirection within the network so that ongoing communications are not interrupted due to the IP address changes of mobile terminals. This is achieved by associating two addresses for the mobile terminal: a permanent home address that represents the terminal's IP address within its home network and a temporary locally assigned care-of address that is valid within the visited network. A home agent (HA) located in the terminal's home network keeps an association between a terminal's home address and its care-of address. A foreign agent (FA) in the visited network provides routing and support services to visiting mobile terminals. The elements of a typical Mobile IP architecture are depicted in Fig. 4.

Both foreign and home agents can advertise their presence on their respective local networks by issuing agent advertisement messages that inform terminals of their availability. Likewise, a mobile terminal may solicit these messages by broadcasting agent solicitation messages on entering a new network to learn about that network's mobility support. A mobile terminal can then use these advertisement messages to detect if it has migrated into a new IP subnet.

Once a mobile terminal learns that it is on a new IP subnet, it will attempt to obtain a care-of address for the visited subnet. This can be done in one of two ways. It may be given a care-of address by the FA, called a *foreign agent care-of address*, which is associated with a network interface on the FA. The mobile terminal may also obtain a *collocated care-of address* associated with one of its own network interfaces. After receiving the care-of address, the mobile terminal will then register this address with its HA via a registration request message. When this registration is accepted, the HA will respond with a registration response message and store the association between the mobile terminal's home address and care-of address.

Packets that are destined for the mobile terminal, that is, those that have the mobile terminal's home address in the destination field of the IP header, always arrive inside the mobile terminal's home network and are intercepted by the HA. The packets are then *tunneled* to the mobile terminal by the HA. The tunneling process involves encapsulating [25] the sender's original IP packet inside the body of another IP packet generated by the HA that contains the mobile terminal's care-of address in the destination field. Since the outer header of the encapsulated IP packet contains the care-of address, it will be forwarded to the visited network where the mobile terminal currently resides. When foreign agent care-of addresses are used, the tunneled packets arrive at the FA who decapsulates them by stripping off the encapsulated packet header and sends the original IP packet to the mobile terminal. Mobile terminals that have collocated care-of addresses will receive incoming encapsulated packets directly and be responsible for decapsulating them.

Bandwidth-constricted wireless networks will most likely support the foreign agent care-of address model, as is the case with the 3GPP and 3GPP2. This mode of operation is more spectrally efficient since only the original IP packet, and not the extra headers associated with the encapsulated packet, is sent over the air interface. Additionally this mode conserves IP address space; one FA can service multiple mobile terminals, and therefore only one care-of address per FA is needed.



Figure 4. Mobile IP architecture.

The mobile terminal can send outbound IP packets on the visited network using normal IP routing without any modifications. The mobile terminal will insert its home address into the source address field of all IP data packets it generates. The mobile IP agents effectively hide mobility events from correspondent hosts so that IP address changes are transparent. Thus correspondent hosts are completely unaware that the mobile terminal has moved. This feature of mobile IP allows all network terminals, regardless of whether they support mobile IP, to communicate with mobile terminals that do.

5.2.1. Route Optimization. Mobility implementations in wireless IP telephony are judged on their ability to provide seamless handoffs with minimal interruption to user sessions. Packet loss and handoff delay therefore must be minimized in order to provide quality voice service. A vulnerability of the basic mobile IP approach discussed above is that traffic streams are required to go to the mobile terminal's HA and then to the mobile terminal, creating what is called the triangular routing problem. This is particularly problematic when a mobile terminal is very far from its home network and is communicating with a correspondent host local to the visited network as shown in Fig. 5. As an example, consider a New Yorker visiting a friend in Los Angeles. Packets from the friend's terminal would have to travel across the country to the HA in New York and then be tunneled back to the present location of the mobile terminal in Los Angeles. This needlessly introduces two cross-country trip times.

In addition, it consumes unnecessary resources within the wide-area network, especially when compared to direct delivery on the local Los Angeles network. Efforts within the mobile IP community have addressed this problem by creating a modified mobile IP approach that uses route optimization [26] and eliminates unnecessary triangular routing.

Route optimization works by making correspondent hosts aware of the current care-of address of the mobile terminal. The host then stores those associations in a binding cache. Home agents, on receipt of a packet destined for a mobile host in a visited network, will send the originator of that packet a message containing the mobile terminal's current care-of address. The originator will then store this association in its binding cache and can begin to send packets directly to the mobile host without unnecessary involvement of the HA. This, in turn, reduces latency and frees network resources. Implementing route optimization in mobile IP can help drastically eliminate latencies and improve quality for wireless IP telephony.

5.3. Mobility Architectures

A design challenge for implementing mobile IP in wireless environments is to define proper placement and demarcations of areas serviced by the mobile IP elements. This helps balance signaling overhead and delay while allowing seamless connectivity. Advertisement and solicitation messages generated in mobile IP are a way of determining whether new physical connections



Figure 5. Mobile IP triangular routing.

require IP-level mobility procedures. However, when cell radii decrease and terminals travel at higher speeds, the mobility rate increases. This triggers more and more solicitation and registration messages that could begin to have a detrimental overhead effect in the network. Furthermore, the registration messages must travel to the mobile terminal's home network and may introduce unacceptable delays in establishing new network connections. Designing wireless networks with subnets that do not cover a lot of area may lead to an undesirable amount of signaling and delay in current mobile IP implementations.

Additionally, mobile agents need to balance the frequency with which they broadcast advertisements to suit the signaling capabilities of the wireless network. More frequent signaling allows for faster detection of network-layer mobility and therefore shorter handoff times. It comes, however, at the price of greater signaling overhead. There exists a design tradeoff that trades signaling overhead for responsiveness of the mobility protocol. As radio-level resources are at a premium, optimal solutions will employ as little signaling overhead as possible to achieve the required level of responsiveness.

Industry efforts have been focused on this problem, and new breeds of mobility strategies have emerged that attempt to reduce the delays and overhead caused by excessive signaling and frequent mobility [27-29]. Many of these strategies introduce levels of hierarchy so that registration messages do not need to travel all the way to the home network every time there is a mobility event. These types of strategies help reduce the latencies and packet losses due to IP mobility.

6. CONCLUSION

Since 1980 wireless voice service has grown into a reliable mainstay for personal and business communications. In the same period the Internet has flourished to unprecedented levels; enjoying economies of scale and ease of deployment. New wireless network architectures will take advantage of the service and management flexibility offered by IP, allowing service providers to offer multimedia and data content to their wireless subscribers. As a consequence, voice strategies must be amended from their traditional circuit-switched roots to perform comparably over IP. The greatest challenges facing the successful deployment of wireless IP telephony are threefold: securing reliable guarantees of service quality on par with traditional cellular systems, obtaining spectral efficiencies over the wireless channel that will not hinder system capacity or service quality, and effectively providing seamless connections to mobile users.

BIOGRAPHY

David Famolari received his B.S and M.S degrees in electrical engineering from Rutgers University, New Jersey, in 1996 and 1999 respectively. In 1996, he joined the Wireless Information Network Laboratory (WINLAB), at Rutgers University, as a research assistant where he worked on radio resource management protocols and parameter optimizations for third generation (3G) cellular systems. Since 1998, he has been a member of the Applied Research Department at Telcordia Technologies, Morristown, New Jersey, where he has worked on emerging mobile computing technologies, wireless networking protocols, and residential networking. David was awarded the Telcordia Technologies CEO Award in 2000 for his contributions in wireless IP networking. He is currently the cochair of the Open Services Gateway Initiative (OSGi) Device Expert Group, a leading industry consortium producing open specifications to promote the delivery of broadband services into home, automotive, and other similar networks. His current research interests include wireless local area network (WLAN) technologies and systems, mobility management, wireless computing, and personal area networks and systems. He can be reached by e-mail at fam@research.telcordia.com.

BIBLIOGRAPHY

- 1. J. Postel, Internetwork Protocol, RFC 791, Sept. 1981.
- H. Holma and A. Toskala, WCDMA for UMTS: Radio Access For Third Generation Mobile Communications, Wiley, New York, 2000.
- 3. TIA/EIA/IS-2000-2-A, *Physical Layer Standard for cdma2000* Spread Spectrum Systems.
- 4. Special Issue, IMT-2000: Standards Efforts of the ITU, *IEEE* Pers. Commun. 4(4): (Aug. 1997).
- S. Blake et al., An Architecture for Differentiated Services, RFC 2475, Dec. 1998.
- V. Jacobson, K. Nichols, and K. Poduri, An Expedited Forwarding PHB, RFC 2598, June 1999.
- 7. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, Assured Forwarding PHB Group, June 1999.
- D. J. Goodman and C. E. Sundberg, Transmission errors and forward error correction in embedded differential PCM, *Bell Syst. Tech. J.* 62: 2735–2764 (Nov. 1983).
- S. H. Lim, D. M. An, and D. Y. Kim, Impact of cell unit interleaving on header error control performance in wireless ATM, *Proc. IEEE GLOBECOM*'96, Nov. 1996, pp. 1705–1709.
- A. Nazer and F. Alajaji, Unequal Error Protection and Source Channel Decoding of CELP Speech Over Very Noisy Channels, Technical Report, Mathematics and Engineering Communications Laboratory, Queens Univ., 1999.
- 11. International Telecommunications Union, One-Way Transmission Time, Recommendation G.114, June 1996.
- C. Comaniciu, N. B. Mandayam, D. Famolari, and P. Agrawal, QoS guarantees for third generation (3G) CDMA systems via admission and flow control, *Proc. IEEE Vehicular Technology Conf. (VTC)*, Boston, Sept. 2000.
- 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects, *QoS Concept and Architecture*, 3GPP TS 23.107 v3.4.0.
- J. Postel, ed., Transmission Control Protocol, RFC-793, Sept. 1981.
- 15. J. Postel, User Datagram Protocol, RFC-768, Aug. 1980.
- H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, Jan. 1996.
- ITU-T, Coding of speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), Recommendation G.729, 1996.
- S. Casner and V. Jacobson, Compressing IP/UDP/RTP Headers for Low-Speed Serial Links, RFC 2508, Jan. 1999.
- M. Degermark, H. Hannu, L. Jonsson, and K. Svanbro, Evaluation of CRTP performance over cellular radio links, *IEEE Pers. Commun.* (Aug. 2000).
- 20. C. Bormann et al., RObust Header Compression (ROHC): Framework and Four Profiles: RTP, UDP, ESP, and Uncompressed, RFC 3095, July 2001.
- L. Jonsson, M. Degermark, H. Hannu, and K. Svanbro, RObust checksum-based header COmpression (ROCCO), Internet draft (June 2000), <draft-ietf-rohc--rtp-rocco-01.txt>.
- 22. H. Hannu, K. Svanbro, and L.-E. Jonsson, ROCCO performance evaluation, Internet draft (May 2000), <draft-ietf-rohc-rtp-rocco-performance-00.txt> (for performance results in ROHC).
- 23. C. Perkins, IP Mobility Support, RFC 2002, Oct. 1996.
- 24. G. Patel and S. Dennet, The 3GPP and 3GPP2 movements toward an all-IP mobile Network, *IEEE Pers. Commun.* (Aug. 2000).
- 25. C. Perkins, IP Encapsulation within IP, RFC 2003, Oct. 1996.
- 26. C. Perkins and D. Johnson, Route optimization in Mobile IP, Internet draft (Sept. 2001), <draft-ietf-mobileip-optim-11.txt>.
- 27. E. Gustafsson, A. Jonsson, and C. Perkins, *Mobile IP* Regional Registration, Internet draft (March 2001), <draftietf-mobileip-reg-tunnel-04.txt>.
- 28. A. Campbell et al., Cellular IP, Internet draft (Dec. 1999), <draft-ietf-mobileip-cellularip-00.txt>.
- 29. R. Ramjee et al., IP micro-mobility support using HAWAII, Internet draft (June 1999), <draft-ietf-mobileip-hawaii-00.txt>.

WIRELESS LAN STANDARDS

RICHARD VAN NEE Woodside Networks Breukelen, The Netherlands

1. INTRODUCTION

Since the early 1990s, wireless local-area networks (WLANs) for the 900-MHz, 2.4-GHz, and 5-GHz ISM (industrial-scientific-medical) bands have been available for a range of proprietary products. In June 1997, the Institute of Electrical and Electronics Engineers approved an international interoperability standard (IEEE 802.11 [1]). The standard specifies both medium-access control (MAC) procedures and three different physical layers (PHY). There are two radio-based PHYs using the 2.4-GHz band. The third PHY uses infrared light. All PHYs support a data rate of 1 Mbps (megabit per second) and optionally 2 Mbps. The 2.4 GHz band is available for license exempt use in Europe, the United States and Japan. Table 1 lists the available frequency

Fable 1. 2	4- and	5-GHz	Bands
-------------------	--------	-------	-------

Location	Regulatory Range (GHz)	Maximum Output Power
North America	2.400 - 2.4835	1000 mW
Europe	2.400 - 2.4835	$100 \text{ mW} (\text{EIRP}^a)$
Japan	2.400 - 2.497	10 mW/MHz
USA (UNII lower band)	5.150 - 5.250	$\begin{array}{l} \text{Minimum of 50 mW or} \\ 4\mathrm{dBm} + 10\log_{10}B^b \end{array}$
USA (UNII middle band)	5.250-5.350	$\begin{array}{l} \text{Minimum of } 250 \text{ mW or} \\ 11 \text{ dBm} + 10 \log_{10} B \end{array}$
USA (UNII upper band)	5.725 - 5.825	$\begin{array}{l} \text{Minimum of 1000 mW or} \\ 17 \text{ dBm} + 10 \log_{10} B \end{array}$

^{*a*}EIRP = effective isotropic radiated power.

^bB is the -26-dB emission bandwidth in MHz.

bands and the restrictions to devices that use this band for communications.

User demand for higher bit rates and the international availability of the 2.4-GHz band has spurred the development of a higher-speed extensions to the 802.11 standard. In 1999, the IEEE 802.11b standard was finished, and describes a PHY providing rates of 5.5 and 11 Mbps [2]. IEEE 802.11b is an extension of the directsequence 802.11 standard, using the same 11 MHz chip rate, such that the same bandwidth and channelization can be used.

In parallel to IEEE 802.11b, the IEEE 802.11a standard was developed to provide high bit rates in the 5-GHz band. This development was motivated by an amendment to Part 15 of the U.S. Federal Communications Commission in January 1997. The amendment made available 300 MHz of spectrum in the 5.2-GHz band, intended for use by a new category of unlicensed equipment called "unlicensed national information infrastructure" (UNII) devices. Table 1 lists the frequency bands and the corresponding power restrictions.

In July 1998, the IEEE 802.11 standardization group decided to select orthogonal frequency-division multiplexing (OFDM) [3] as the basis for their new 5-GHz standard, targeting a range of data rates from 6 to 54 Mbps. This standard is the first one to use OFDM in packet-based communications, while the use of OFDM previously was limited to continuous transmission systems like digital audiobroadcasting (DAB) and digital videobroadcasting (DVB). Following the IEEE 802.11 decision, the European HIPERLAN type 2 [4] standard and the Japanese Multimedia Mobile Access Communication (MMAC) standard also adopted OFDM. The three bodies have worked in close cooperation since then to minimize differences between the various standards, thereby enabling the manufacturing of equipment that can be used worldwide.

Regulatory issues played an important role in the development of wireless LAN standards. One of the key factors in the choice of modulation schemes for the 2.4-GHz band has been the FCC spreading requirement for unlicensed devices in the ISM bands, where wireless LANs

are predominantly used. According to the FCC spreading rules, transmission in the ISM bands have to use either direct sequence, spread spectrum, or frequency hopping. Frequency-hopping devices have to use at least 75 hopping channels with a maximum dwell time of 400 ms. Directsequence devices have to demonstrate at least 10 dB processing gain in a narrowband jammer test, which basically shows that there is a gap of at least 10 dB between signal-to-noise ratio and signal-to-interference ratio requirements for a certain bit error ratio. In the early days of wireless LAN, many people interpreted the spreading rule as a requirement for at least 10 chips per symbol; hence the 11 chips spreading sequence in the 802.11 standard. Later, a less strict interpretation was adopted, purely based on meeting the narrowband jammer test. This is clearly visible in the IEEE 802.11b standard. The 802.11b standard uses complementary code keying (CCK), which can be viewed as direct-sequence spreadspectrum modulation with multiple spreading codes with a length of 8 chips. Despite the less strict interpretation, the spreading rule formed a barrier for really high data rates. It blocked the use, for instance, of OFDM in the 2.4-GHz band. In order not to avoid further technological progress in the 2.4-GHz band, in May 2001 the FCC decided to allow digital transmissions without any spreading requirement [5]. This opened the way to higher data rates using OFDM in the 2.4-GHz band. The 802.11 committee took advantage of this rule change by selecting the OFDM based 802.11a standard as basis for the 802.11g standard, extending the data rates in the 2.4-GHz band up to 54 Mbps.

In the following sections we describe the various IEEE wireless LAN standards, and mention the differences with HIPERLAN and MMAC. Because of length limitations, the scope of this article is restricted to the most predominantly used parts of the standards. More details can be found in the references listed at the end of this article.

2. IEEE 802.11 MAC

The IEEE 802.11 MAC standard consists of one mandatory and two optional modes [1]. All modes use time-division duplex (TDD), so the medium is shared in time between different users and/or access points. The mandatory part is the *distributed coordination function* (DCF), which uses carrier sense multiple access with collision avoidance (CSMA/CA). Figure 1 shows the timing diagram of a DCF packet transmission. Before starting a transmission, the channel is sensed to see if it is available. If no other signal is received above a certain defer threshold, a packet is send. After successful reception, the recipient sends an acknowledgment back. After receiving the acknowledgement, the first user has to wait for a time DIFS plus a random backoff time before transmitting another packet. DIFS is the *distributed interframe spacing*, which is equal to the *short interframe spacing* between packet and acknowledgment plus 2 slot times.

Optional modes in the 802.11 MAC are the requestto-send/clear-to-send (RTS/CTS) protocol, and the point coordination function (PCF). PCF is a centralized MAC, where an access point polls stations to see if they have packets to transmit. PCF can be used to guarantee a minimum packet delay, but it can do this only in the absence of interference from other cells. With the RTS/CTS protocol, prior to a data packet, first a short request packet is sent. The receiver answers with a CTS packet, which contains a net allocation vector (NAV) that tells all users how long the current RTS/CTS cycle will take. The effect of this is that all users that can receive the CTS packet will not try to try to compete for the channel for the duration indicated by the NAV. This solves the hiddennode problem of DCF without RTS/CTS, because in that case, only users that can receive the transmitter will stop competing for the channel. So, it can happen that a packet from user A to B is interfered by user C, who does not have a good link to A, but it does have a good link to B. With RTS/CTS, this situation is avoided, because user Cwill hear the CTS coming from *B*.

3. IEEE 802.11 DSSS

The IEEE 802.11 Direct-Sequence Spread-Spectrum standard is based on the transmission of 11-chip Barker codes at a 11 MHz chip rate. Data rates of 1 and 2 Mbps are achieved using BPSK or QPSK modulation of the Barker codes, respectively. The 11-chip Barker code is defined as $\{1, -1, 1, 1, -1, 1, 1, 1, -1, -1, -1\}$. Its primary use is to satisfy the FCC spreading requirements, as well as providing robustness against multipath propagation and narrowband interferers. Robustness against multipath is obtained by the ideal aperiodic autocorrelation properties that define a Barker code - a Barker code is a code for which the absolute autocorrelation sidelobes are equal to or less than one (≤ 1) for all nonzero delays, compared to L for a zero delay, where L is the codelength. Because of the low-autocorrelation sidelobes, effects of intersymbol interference are greatly suppressed, while a simple RAKE receiver is able to significantly benefit from multipath diversity in frequency selective channels.

The 802.11 packet structure is shown in Fig. 2. The complete packet (PPDU) has three segments. The first segment is the preamble, which is used for signal detection and sychronization. The second segment is the header, which contains data rate and packet length information. The third segment (MPDU) contains the information bits.



Figure 1. Timing diagram of a single-packet transmission using DCF.



Figure 2. The packet structure used for 802.11 DSSS 1 and 2 Mbps, with the extension to 5.5 and 11 Mbps shown.

The preamble and header are transmitted at 1 Mbps, while the data portion is sent at one out of four possible rates.

The preamble is formed from a SYNC field and a SYNC field delimiter (SFD). The SYNC field is generated using 128 scrambled ones. The SYNC field is used for clear channel assessment, signal detection, timing acquisition, frequency acquisition, multipath estimation, and descrambler synchronization.

4. IEEE 802.11b

In July 1998, the IEEE 802.11b working group adopted complementary code keying (CCK) as the basis for the high-rate physical-layer extension to deliver data rates of 5.5 and 11 Mbps [6]. This high-rate extension was adopted in part because it provided an easy path for interoperability with the existing 1- and 2-Mbps networks by maintaining the same bandwidth and utilizing the same preamble and header as shown in Fig. 1. An optional short preamble with a 56-bit SYNC field is specified to increase the net data throughput.

Complementary codes were originally conceived by M. J. E. Golay for infrared multislit spectrometry [7]. However, their properties also make them useful in radar applications and more recently for discrete multitone communications and OFDM [8]. The original publication [7] defines a complementary series as a pair of equally long sequences composed of two types of elements that have the property that the number of pairs of like elements with any given separation in one series is equal to the number of pairs of unlike elements with the same separation in the other series. Another way to define a pair of complementary codes is to say that the sum of their aperiodic autocorrelation functions is zero for all delays except for a zero delay.

The CCK codes that were selected as the basis for IEEE 802.11b were first published in 1996 [8]. More background information on these codes can be found in Halford et al. [9]. The following equation represents the 8

complex chip values for the CCK code set, with the phase variables being QPSK phases:

$$c = \{ e^{j(\varphi_1 + \varphi_2 + \varphi_3)}, e^{j(\varphi_1 + \varphi_3 + \varphi_4)}, e^{j(\varphi_1 + \varphi_2 + \varphi_4)}, -e^{j(\varphi_1 + \varphi_2 + \varphi_4)}, e^{j(\varphi_1 + \varphi_2)}, e^{j(\varphi_1 + \varphi_2)$$

Basically, the three phases φ_2 , φ_3 and φ_4 , define 64 different codes of 8 chips, where φ_1 gives an extra phase rotation to the entire codeword. Actually, the latter phase is differentially encoded across successive codewords, equivalent to the 1- and 2-Mbps DSSS differential phase encoding. This feature allows the receiver to use differential phase decoding, eliminating a carrier tracking PLL, if desired. Each of the four phases φ_1 to φ_4 represents 2 bits of information, so a total of 8 bits is encoded per 8-chip CCK codeword.

At 5.5 Mbps, the processing is similar. Four information bits are consumed per 8-chip CCK codeword transmission. The codeword rate is still 1.375 MHz, since the chip rate is 11 Mchips/s. Two bits select 1-of-4 CCK subcodes. The other two information bits quadriphase-modulate (rotate) the whole codeword. The 4 CCK subcodes are contained in the larger 64 subcode set of 11 Mbps. At the receiver, the CCK codes can be decoded by using a modified fast Walsh transform as described by Grant and van Nee [10].

5. IEEE 802.11a

IEEE 802.11a provides data rates of 6–54 Mbps in the 5-GHz band using orthogonal frequency-division multiplexing (OFDM). The basic principle of OFDM is to split a high-rate datastream into a number of lowerrate streams that are transmitted simultaneously over a number of subcarriers. Since the symbol duration increases for the lower-rate parallel subcarriers, the relative amount of time dispersion caused by multipath delay spread is decreased. Intersymbol interference is eliminated almost completely by introducing a guard time in every OFDM symbol. In the guard time, the OFDM symbol is cyclically extended to avoid intercarrier interference. Figure 3 shows an example of 4 subcarriers



Figure 3. OFDM symbol with cyclic extension.

from one OFDM symbol. In practice, the most efficient way to generate the sum of a large number of subcarriers is by using inverse fast fourier transform (IFFT). At the receiver side, FFT can be used to demodulate all subcarriers. It can be seen in Fig. 3 that all subcarriers differ by an integer number of cycles within the FFT integration time, which ensures the orthogonality between the different subcarriers. This orthogonality is maintained in the presence of multipath delay spread, as illustrated by Fig. 3. Because of multipath, the receiver sees a summation of time-shifted replicas of each OFDM symbol. As long as the delay spread is smaller than the guard time, there is no intersymbol interference nor intercarrier interference within the FFT interval of an OFDM symbol. The only remaining effect of multipath is a random phase and amplitude of each subcarrier, which has to be estimated in order to do coherent detection. In order to deal with weak subcarriers in deep fades, forward error correction across the subcarriers is applied.

5.1. OFDM Parameters

Table 2 lists the main parameters of the IEEE 802.11a OFDM standard. A key parameter that largely determined the choice of the other parameters is the guard interval of 800 ns. This guard interval provides robustness to rootmean-squared delay spreads up to several hundreds of nanoseconds, depending on the coding rate and modulation used. In practice, this means that the modulation is robust enough to be used in any indoor environment, including large factory buildings. It can also be used in outdoor environments, although directional antennas may be needed in this case to reduce the delay spread to an acceptable amount and to increase the range.

In order to limit the relative amount of power and time spent on the guard time to 1 dB, the symbol duration was chosen to be $4 \mu s$. This also determined the subcarrier spacing to be 312.5 kHz, which is the inverse of the symbol duration minus the guard time. By using 48 data subcarriers, uncoded data rates of 12-72 Mbps can be

Table 2. Main Parameters of the OFDM Standard

Data Rate	6, 9, 12, 18, 24, 36, 48, 54 Mbps
Modulation	BPSK, QPSK, 16-QAM, 64-QAM
Coding rate	$\frac{1}{2}, \frac{2}{3}, \frac{1}{3}$
Number of subcarriers	$\frac{1}{52}$
Number of pilots	4
OFDM symbol duration	4 μs
Guard interval	800 ns
Subcarrier spacing	312.5 kHz
-3-dB bandwidth	16.56 MHz
Channel spacing	20 MHz

achieved by using variable modulation types from BPSK to 64-QAM. In addition to the 48 data subcarriers, each OFDM symbol contains an additional 4 pilot subcarriers, which can be used to track the residual carrier frequency offset that remains after an initial frequency correction during the training phase of the packet.

In order to correct for subcarriers in deep fades, forward error correction across the subcarriers is used with variable coding rates, giving coded data rates of 6–54 Mbps. Convolutional coding is used with the industry standard rate- $\frac{1}{2}$, constraint length 7 code with generator polynomials (133,171). Higher coding rates of $\frac{2}{3}$ and $\frac{3}{4}$ are obtained by puncturing the rate- $\frac{1}{2}$ code.

5.2. Channelization

For the 200-MHz-wide spectrum in the lower and middle UNII bands, 8 OFDM channels are available with a channel spacing of 20 MHz. The outermost channels are spaced 30 MHz from the band edges in order to meet the stringent FCC-restricted band spectral density requirements. The FCC also defined an upper UNII band from 5.725 to 5.825 GHz, which carries another 4 OFDM channels. For this upper band, the guard spacing from the band edges is only 20 MHz, since the out-of-band spectral requirements for the upper band are less severe than those of the lower and middle UNII bands. In Europe, the same spectrum as the lower and middle UNII band is available, plus an extra band from 5.470 to 5.725 GHz. In Japan, a 100-MHz-wide band from 5.15 to 5.25 is available. This band contains 4 OFDM channels with 20 MHz guard spacings from both band edges.

5.3. OFDM Signal Processing

The general block diagram of the baseband processing of an OFDM transceiver is shown in Fig. 4. In the transmitter path, binary input data are encoded by a standard rate- $\frac{1}{2}$ convolutional encoder. The rate may be increased to $\frac{2}{3}$ or $\frac{3}{4}$ by puncturing the coded output bits. After interleaving, the binary values are converted into QAM values. To facilitate coherent reception, 4 pilot values are added to each 48 data values, so a total of 52 QAM values is reached per OFDM symbol, which are modulated onto 52 subcarriers by applying the inverse fast Fourier transform (IFFT). To make the system robust to multipath propagation, a cyclic prefix is added. Further, windowing is applied to get a narrower output spectrum. After this step, the digital output signals can be converted to analog signals, which are then upconverted to the 5 GHz band, amplified, and transmitted through an antenna.

The OFDM receiver basically performs the reverse operations of the transmitter, together with additional training tasks. First, the receiver has to estimate frequency offset and symbol timing, using special training symbols in the preamble. Then, it can do a FFT for every symbol to recover the 52 QAM values of all subcarriers. The training symbols and pilot subcarriers are used to correct for the channel response as well as remaining phase drift. The QAM values are then demapped into binary values, after which a Viterbi decoder can decode the information bits.

Figure 5 shows the time-frequency structure of an OFDM packet, where all known training values are marked in gray. It illustrates how the packet starts with 10 short training symbols, using only 12 subcarriers, followed by a long training symbol and data symbols, with each data symbol containing 4 known pilot subcarriers that are used

for estimating the reference phase. The preamble, which is contained in the first 16 μ s of each packet, is essential to perform start-of-packet detection, automatic gain control, symbol timing, frequency estimation, and channel estimation. All of these training tasks have to be performed before the actual data bits can be successfully decoded. More detailed information on OFDM signal processing as well as performance results can be found in Ref. 11.

5.4. Differences Between IEEE, ETSI, and MMAC

The main differences between IEEE 802.11 and HIPER-LAN type 2—which is standardized by ETSI BRAN—are in the medium access control (MAC). IEEE 802.11 uses a distributed MAC based on carrier sense multiple access with collision avoidance (CSMA/CA), while HIPERLAN type 2 uses a centralized and scheduled MAC, based on wireless ATM. MMAC supports both of these MACs. As far as the physical layer is concerned, there are only a few minor differences, summarized as follows:

- HIPERLAN uses extra puncturing to accommodate the tail bits in order to keep an integer number of OFDM symbols in 54-byte packets [12].
- In the case of 16-QAM, HIPERLAN uses rate $\frac{9}{16}$ instead of rate $\frac{1}{2}$ —giving a bit rate of 27 instead of 24 Mbps—in order to get an integer number of OFDM symbols for packets of 54 bytes. The rate $\frac{9}{16}$ is made by puncturing 2 out of every 18 coded bits.
- HIPERLAN uses different training sequences. The long training symbol is the same as for IEEE 802.11, but the preceding sequence of short training symbols is different. A downlink transmission starts with 10 short symbols as in IEEE 802.11, but the first 5 symbols are different in order to enable detection of the start of the downlink frame. Uplink packets may use 5 or 10 identical short symbols, with the last short symbol inverted.

6. IEEE 802.11g

The IEEE 802.11g standard extends the 802.11b standard with higher data rates for the 2.4-GHz band [13]. It



Figure 4. Block diagram of OFDM transceiver.



Figure 5. Time-frequency structure of an OFDM packet. Gray-shaded subcarriers contain known training values.

achieves this by simply allowing IEEE 802.11a OFDM transmissions in the 2.4-GHz band, while 802.11a was originally targeted at the 5-GHz band. To remain coexistent with legacy 802.11b devices, the 802.11 RTS-CTS mechanism can be used to claim airtime for high-rate OFDM packets. If no legacy devices are present in a network, it is possible to transmit 802.11a packets without the RTS-CTS mechanism to maximize user throughput. The IEEE 802.11g standard also defines two optional modulation schemes. The first is CCK-OFDM, where each OFDM packet is preceded by an 802.11b header to provide full coexistence with legacy 802.11b devices without the need for RTS-CTS. The second option is packet binary convolutional coding (PBCC), which provides a 22-Mbps raw data rate using coded 8-PSK together with a standard 802.11b header.

With the advent of IEEE802.11g, OFDM has become the single solution for high data rates in both the 2.4- and 5-GHz bands, thereby facilitating the production of dualband devices. While OFDM is ideal for high data rates, it is expected that the old 1- and 2-Mbps 802.11 rates in the 2.4-GHz band will remain important for providing the largest possible coverage range.

BIOGRAPHY

Richard van Nee before his employment as director of WLAN Product Engineering at Woodside Networks, Dr. Van Nee was a key member of the technical staff at Lucent Technologies/Bell Labs in the Netherlands. Dr. Van Nee was among those who proposed the OFDM-based physical layer, which was selected for standardization in IEEE 802.11, MMAC, and ETSI HiperLAN. He was involved in the design of the OFDM modems for the European Magic WAND project. Together with NTT, he made the original OFDM-based proposal that lead to the IEEE 802.11a wireless LAN high-rate extension for the 5 GHz band, with data rates up to 54 Mbps. He also was one of the original proposers of the 11 Mbps IEEE 802.11b extension for the 2.4 GHz band, which is based on Complementary Code Keying as described in one of his papers from 1996. Together with Prof. Ramjee Prasad from Aalborg University, Denmark, he wrote the book OFDM for Mobile Multimedia Communications, which is a well-read reference for anyone involved with the new IEEE 802.11a standard. He received his Ph.D. in electrical engineering from Delft University, and his M.Sc. in electrical engineering from Twente University.

BIBLIOGRAPHY

- 1. IEEE, 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Nov. 1997.
- IEEE 802.11, Supplement to IEEE Standard for Information Technology—Telecommunications and Information Exchange between Systems—LAN/MAN Specific Requirements—Part 11: Wireless MAC and PHY Specifications: Higher Speed Physical Layer in the 2.4 GHz Band, IEEE Standard 802.11b, Jan. 2000.
- IEEE 802.11, Supplement to IEEE Standard for Information Technology—Telecommunications and Information Exchange between Systems—LAN/MAN Specific Requirements—Part 11: Wireless MAC and PHY Specifications: High Speed Physical Layer in the 5 GHz Band, IEEE Standard 802.11a, Dec. 1999.

- 4. ETSI, Radio Equipment and Systems, HIgh PErformance Radio Local Area Network (HIPERLAN) Type 1, European Telecommunication Standard, ETS 300-652, Oct. 1996.
- Federal Communications Commission Notice of Proposed Rulemaking and Order, FCC 01-158, ET Docket 99-231, May 11, 2001.
- M. Webster, C. Andren, J. Boer, and R. van Nee, Harris/Lucent TGb Compromise CCK (11 Mbps) Proposal, Document IEEE P802.11-98/246, July 1998.
- M. J. E. Golay, Complementary series, *IRE Trans. Inform.* Theory 82–87 (April 1961).
- R. van Nee, OFDM codes for peak-to-average power reduction and error correction, *IEEE Global Telecommunications Conf.* Nov. 18-22, 1996, pp. 740-744.
- 9. K. Halford et al., Complementary Code Keying for Rake-Based Indoor Wireless Communication, IEEE ISCAS '99, Orlando, FL.
- A. Grant and R. van Nee, Efficient maximum likelihood decoding of Q-ary modulated Reed-Muller codes, *IEEE* Commun. Lett. 2(5): 134-136 (May 1998).
- 11. R. van Nee and R. Prasad, OFDM for Mobile Multimedia Communications, Artech House, Boston, Jan. 2000.
- 12. ETSI BRAN, HIPERLAN Type 2 Functional Specification Part 1—Physical Layer, DTS/BRAN030003-1, June 1999.
- S. Halford et al., Proposed Draft Text: B + A = G High Rate Extension to the 802.11b Standard, draft IEEE 802.11G Standard, Document IEEE 802.11-01/644r0, November 2001.

WIRELESS LOCAL LOOP STANDARDS AND SYSTEMS

Homayoun Hashemi Sharif University of Technology Teheran, Iran

1. INTRODUCTION

Communication plays a vital role in economic development of nations, and in prosperity and well-being of their citizens. A communication plant consists of three main segments: (1) local access plant, the "last mile" of communication, where a subscriber (home or office) is connected to the telephone company's central office; (2) switching facilities, the mechanism that switches and routes calls to their final destination; and (3) long-distance transmission lines, through which calls are transferred within remote areas of the same nation, or between different countries. Segments 2 and 3 have undergone drastic changes in the >125-year history of telephony. The manual switches of early era were transformed into the more advanced electromechanical switches, which then evolved to today's high-speed electronic (digital) switches. "Long distance" lines of 100 years ago consisted primarily of copper wires installed on tens of thousands of telephone poles between cities. Today, millions of calls are transferred nationally and internationally every day by vast and complicated interconnection of high-capacity microwave lines, fiberoptic networks, and low- and highorbit satellites. The local access technology, however,

remains fundamentally unchanged since the times of Alexander Graham Bell. The dominant local loop up to the mid 1990s was a twisted pair of copper line buried underground to connect the home to the nearest telephone exchange.

Telecommunication systems have gone through three phases of evolution [1]. In the era of interconnection (1876-1950) homes and businesses across cities were wired up to telephone central offices (COs), and COs within a city were connected by wire. Limited long-distance lines were established between cities and nations, again, dominantly by copper wire. In the era of networks (1950–1990) the core network was transformed by expanding, automating, and expediting the switching, call processing, and transport functions. Explosive growth of telephony in this era was fueled by invention and successful deployment of equipment and systems such as digital computers, digital switches, satellites, fiberoptics, Integrated Services Digital Network (ISDN), asynchronous transfer mode (ATM) switches, and the Internet. These inventions revolutionized telecommunications, and drastically changed the way people live and work. Great expansion of long-distance telephony, automatic dialing, and introduction of intelligent networking functions are among the achievements of this era. As an example, only 1% of the "core network" in the United States in 1990 was developed in the first 75 years of "interconnection era." The balance of 99% was developed in the 40 years of "network era" [1].

The invention and rapid deployment of cellular radio systems in the 1980s has paved the way for a new era in telecommunications, the "era of access," which started roughly in 1990. This era is manifested by gradual replacement of the primitive, costly, and difficult to install and maintain wired (copper)-based local loops, to the modern, relatively inexpensive, and easy to install wireless local loops (WLLs). It is anticipated that in the near future the majority of telephone services in the world will be based on wireless access (fixed or mobile). Replacement of wired local loop in a sense removes the bottleneck of communications, paves the way for quick, efficient, and economical introduction of other services such as data and video, in addition to the traditional speech communications, a process that expedites transformation of our society into the Information Age.

Basic principles of WLL, including its technical, economical, and regulatory aspects, are described in the literature [1-6].

2. WLL PRINCIPLES

A wireless access system, which is also known as WLL, fixed cellular radio, fixed wireless access (FWA), radio in the local loop (RLL), is emerging as a modern access method. Basic wireline and wireless access system models are shown in Fig. 1. Inspection of this figure shows that a multiple access radio system replaces the wires in the loop.

WLL systems consist of four basic building blocks: (1) a "radio terminal" or "subscriber station," the visible transceiver device carried by or located near the user; (2) a "radio base station," which provides the "wireless" air interface between the subscriber and the network;



Figure 1. Basic wireline and wireless access system models. (*Source*: Ref. 2.)

 $(3)\,a$ "network control subsystem," which controls the wireless access system; and $(4)\,a$ "fixed network," the wired infrastructure to which the wireless system provides access.

This scenario can be implemented at each location, independent of other locations, to provide access to limited number of subscribers. The large-scale implementation in an area, however, requires application of spectrumefficient techniques. Modern WLL is therefore based on principles of cellular radio.

2.1. The Cellular Radio Concept

Cellular radio has been developed for mobile telecommunications, out of a need to increase system capacity and at the same time conserve the scare radio spectrum. The concept is very simple; a large geographic area is partitioned into smaller local areas called "cells." The block of radio spectrum (normally consisting of hundreds of channels) is also partitioned into smaller subblocks or "sets." Channel sets are then assigned to each cell. Two distant cells can use the same channel sets without excessive interference with each other. Repeated reuse of the same channel sets throughout a service area results in drastic increases in system capacity [7]. A mobile subscriber initializes a call while moving in a given cell. When he crosses the cell boundary, calls are "handed off" to the new base station serving the new cell. This process of changing carrier frequency is performed automatically, without any subscriber action.

When telephone traffic in a service area increases, the system capacity can be increased accordingly by "cell splitting," where new fixed antennas or base stations are placed half-way between existing antennas, splitting large calls into smaller cells. This process increases frequency reuse and therefore system capacity. The cell splitting process is gradual and nonuniform to reflect the nonuniformity of telephone traffic and differences in growth rate throughout a service area.

Cellular radio, originally developed for providing voice communication to mobile (vehicular) subscribers, has evolved since the early 1980s into a sophisticated wireless engine capable of providing variety of services to both fixed and mobile users. Basic principles are described by Lee [8] and Rappaport [9].

2.2. Differences Between Fixed and Mobile Access

Application of cellular radio to WLL is similar to mobile radio, with the exception that in WLL both ends of transmission are normally fixed. This brings a number of advantages [5]:

- 1. The handoff procedure is not implemented, resulting in simplifications of system design and resource management, and in reduction of system controller's processing capability requirements. It should be noted that even if the user moves around moderately, this advantage still applies since the coverage area (or cell) does not change.
- 2. Both antennas are higher, resulting in a line-of-sight link most of the time. Propagation loss is, therefore, smaller, and coverage area is larger.
- 3. Fixed subscriber transceiver makes higher transmission powers (as compared to mobile) possible.
- 4. Directional antennas can be implemented at subscriber site, as well as at the base station. This results in higher gain transmission to the desired location, and limited interference to other sites using the same frequency. Smaller overall interference increases frequency reuse and system capacity.
- 5. Properties 2–4 (above) result in an increase in coverage area. This reduces number of required cells in an area, a great advantage for low-density sparsely populated rural locations where system capacity is not an issue.
- 6. Unlike mobile access, the nature of traffic is not dynamic. This makes frequency planning easier and more efficient.
- 7. Stationarity of the subscriber unit results in the absence of short-term multipath fading channel impairments. This results in better quality of service.

In spite of a number of differences between WLL and mobile wireless access, since both are based on the cellular concept, a number of standards originally developed for mobile subscribers have been successfully used for WLL applications.

2.3. WLL Subscriber Base and Types of Service

WLL can provide service to both developed industrial nations, and developing countries.

In developed countries, where a reasonably extensive wireline-based communication infrastructure is in place and telephone penetration is high, WLL can be used to provide cost-effective and easy-to-implement service to low-density remote areas. It can also provide additional low-cost, quick-to-implement service to high-density urban areas.

In developing countries where telephone penetration in normally very low, building a copper-based infrastructure is very expensive and time-consuming. WLL can provide a suitable answer to the basic needs of developing countries by injecting hundreds of thousands of lines into each metropolitan area in a short span of time. Rapid improvements in communication facilities of these countries improve standards of living and reduce economic gaps with developed nations [5]. Cellular radio layout with large cells can also be implemented in remote areas of developing countries to provide single-line service to each village lacking telecommunication privileges. This is the fastest and cheapest method to connect the rural population to the communication network.

It should be noted that communication requirements of developing and developed countries are different. Developing countries mostly need voice telephone lines. Service requirement emphasizes low cost and high capacity, even if voice quality is to be sacrificed. In developed industrial countries emphasis is on quality of service and capability to provide new services. Although WLL principles are the same for both, system design aspects are different [10,11].

Low-capacity WLL systems, not based on cellular, have been used for many decades to provide telecommunication (voice) services to isolated and remote areas of both developed and developing countries. More recently, however, new communication services such as fax, data, and video have grown tremendously. The explosive growth of the Internet technology and services has resulted in an increase in data transmission requirements of modern offices and homes. Such services, a number of which are broadband in nature, have changed telecommunication needs of the subscribers, and have brought new requirements on telecommunication facilities and infrastructure. WLL is also capable of providing these new services. For broadband applications, however, availability of spectrum is an issue. Design concepts are also different.

2.4. Advantages of WLL

Deployment of WLL is gaining momentum for providing service to densely populated urban areas, as well as sparsely-populated remote and rural areas [2–5]. Replacing the wires in the "last mile" of communication with WLL results in the following major advantages:

- Speed of Implementation. Installation of wired local loops involves burial of wires underground, a timeconsuming process, especially for longer paths. In WLL, on the other hand, radio units can be installed quickly at both ends of transmission. The speed of implementation within the radio coverage area is insensitive to distance. WLL can be implemented 5-10 times faster than copper-based systems [2].
- 2. Small Initial Investment. Wired local loops require large lumpy investments, as shown in Fig. 2. A



Figure 2. Costs of wireline and wireless access systems versus time. (Source: Ref. 1.)

2950 WIRELESS LOCAL LOOP STANDARDS AND SYSTEMS

large initial investment is required to lay down main distribution cables, a capacity that remains unused for long periods of time. Subsequent major expansions also involve other lumps of added investments, which bring no immediate return. WLL, on the other hand, requires investments in small increments to match growth in traffic (Fig. 2). Roughly 20% of installation costs are related to the infrastructure; the remaining 80% is spent at the time when subscriber receives service, which is followed by revenues and immediate return on investment. These advantages make competition feasible for small startup companies with limited capital.

- 3. Cheap and Easy Maintenance. Operational costs of WLL are considerably less than those of wireline loops. One study shows a reduction of 25% per subscriber per year [2]. This study shows that over 30% of trouble reports for wired networks are related to distribution cable, drop wires, and in-home wiring, which all can be eliminated in WLL-based systems. Theft and vandalism of wired loops are other types of loss for telephone companies, which can be avoided by WLL.
- 4. Fast and Easy Substitution of Faulty Equipment. Telecommunication facilities in WLL-based systems are located either at the central office (CO) or in customer premises. In wired loops underground copper wires substitute a major portion of the hardware. Repairs and substitutions in WLL are therefore much faster.
- 5. *Possibility of Deinstallation and Redeployment*. If subscriber demand in an environment declines, wired loops are simply abandoned, with great capital losses (Fig. 2). In WLL, on the other hand, equipment consisting of radio units at both ends of transmission can be removed and redeployed in other places.
- 6. Insensitivity to Subscriber's Exact Location. Implementation of copper loops requires knowledge of exact subscriber location, in contrast with WLL, in which the subscriber only needs to be within the radio coverage area.
- 7. *Mobility of Subscriber*. The emphasis of WLL is on providing service to fixed terminals. The radio loop, however, results in the added advantage of subscriber mobility.
- 8. Variety of Services. The twisted pair of copper is a primitive, very-low-capacity transmission medium that has been used traditionally for single-channel voice communication. Effective line capacity of wired loops has been increased by introducing sophisticated (and relatively expensive) digital subscriber loop (DSL) modems at both ends of the loop. Still, the wired loop provides a bottleneck for delivery of broadband video and high bit rate data to subscribers. In radio loops, however, such broadband services can be offered by deploying line-of-sight (LoS) transmission at higher frequencies. Availability of spectrum is, of course, an issue in WLL.

3. WLL TECHNOLOGY

Since large-scale deployment of WLL is based on cellular radio, most techniques and technologies developed for mobile radio applications can be successfully applied to WLL.

3.1. Access Methods

Three major channel access methods for cellular wireless communications are frequency-division multiple access (FDMA), time-division multiple access (TDMA), and codedivision multiple access (CDMA) [8,9].

Under an FDMA scheme the allocated band is divided into a number of distinct channels, each one to be used as a single-voice channel. Signals are therefore separate in frequency but mixed in time. In FDMA, when all channels in a cell are occupied, a new call is blocked. FDMA with frequency modulated voice was applied successfully to large-scale mobile telephony in the 1980s. Typically 500–1000 narrowband duplex channels were allocated to a system.

In TDMA, which should better be labeled FDMA/TDMA, the allocated band is first divided into a number of distinct physical channels. Unlike FDMA, however, each physical channel is now used for time multiplexing of several messages. Each user is assigned one of a number of nonoverlapping time slots during which he/she can send or receive digitized messages. Signals are therefore separate in time but mixed in frequency. Transmitter and receiver should have precise time synchronization to avoid interchannel crosstalk. A number of mobile cellular standards operating on the FDMA/TDMA principle were developed in the 1980s and implemented in the 1990s.

CDMA is a spread-spectrum technique, and therefore benefits from associated antinoise, antiinterference capabilities. In CDMA signals are mixed both in time and frequency. Each user in a cell is assigned a distinct code that has a large bandwidth. Codes have good orthogonal properties. At the receiver a correlator is used to correlate the received signal with a replica of the transmitted code (for that particular user). The original message is therefore reconstructed. One great advantage of CDMA is its high capacity. In cellular CDMA, which is also FDMA/CDMA in nature, every physical channel is assigned to each cell. This process eliminates the need for frequency planning, in addition to increasing capacity. CDMA enjoys a "soft capacity limit," in which an additional user can always be accommodated at the expense of small added interference to all other users in the same cell. Successful implementation of CDMA requires accurate synchronization and appropriate power control capabilities. Performance and capacity of CDMA for mobile radio applications were subjects of intense debate in the 1990s. Only one major mobile radio standard based on CDMA emerged in the 1990s. More recently developed wireless communication standards are, however, based mostly on CDMA technology.

TDMA and CDMA can each be either narrowband or wideband. This refers to the width of each physical channel. Wideband schemes have higher capacities and are capable of accommodating new high-bit-rate nonvoice services. Implementation, however, is more difficult. Newly emerging wireless standards are widebandoriented.

WLL systems can operate in both frequency-division duplex (FDD) and time-division duplex (TDD) modes. In FDD a pair of duplex channels, widely separated in frequency, are assigned to a user for two-way transmission. In TDD the same physical channel is used for both downlink (base-to-subscriber) and uplink (subscriber-tobase) transmissions. Each subscriber using the physical channel, however, sends and receives messages at different time slots. With exceptions, most currently working wireless systems operate on FDD mode. The emerging systems use both FDD and TDD.

3.2. Resource Management

Successful deployment of large-scale WLL systems depends on efficient use of scarce radio spectrum, which in turn requires good channel assignment policies. There are two major techniques: fixed channel assignment (FCA), and dynamic channel assignment (DCA). Each technique contains a number of variations, and a combination of the two has also been suggested.

In FCA the available channels are partitioned into blocks or "sets." During initial planning of the system, channel sets are assigned to each cell according to forecast traffic requirements. Although initial planning of FCA is sophisticated and requires knowledge of traffic distributions throughout the service area, its later operation is relatively simple. The great disadvantage of FCA is its lower capacity (compared to that of DCA). This is particularly true where traffic distribution is nonuniform.

In DCA all channels are assigned to every cell. Channel assignment for each call is performed by a central processor on an individual basis, after taking interference limitation requirements of the system into account. The advantage of DCA is minimal initial planning and high capacity, especially where traffic distribution is nonuniform in space, and changing with time. The major disadvantage of DCA is elaborate call supervision, which puts a heavy burden on the central processor for small-cell high-capacity systems. With exceptions, current mobile wireless systems use FCA. The emerging standards, however, take advantage mostly of DCA.

All currently available wireless (fixed or mobile) systems are based on circuit switching, where a dedicated circuit is allocated to a user throughout the connection. The circuit may be an FDMA channel, a time slot of a TDMA channel, or a CDMA orthogonal code. On termination of the call the circuit is marked "idle," and later assigned to a new user. The newly emerging standards operate on both circuit-switching and packet-switching modes. In packet switching there is no permanent connection for a user. A number of calls are made using the same channel. This channel is assigned to a number of users on a temporary basis. Each user sends its information in "packets" at assigned intervals. This scheme increases efficiency, and hence capacity, but requires great call supervision and sophisticated processing. Although packet switching can be used for speech, it is more suitable for data transmission applications, which are bursty in nature.

4. WLL STANDARDS

In principle any wireless personal communication standard can be used for WLL systems. The standards, however, are grouped in two major categories: standards based on existing and emerging digital mobile radio systems, and those based on proprietary radio technologies. The first category covers open standards developed by recognized standardization bodies. They provide network operators with the freedom of supplying different subsystems from different manufacturers. In the second category the entire system is normally developed by a specific manufacturer based on proprietary-developed standards.

4.1. Mobile-Radio-Based Standards

Wireless mobile communication is the fastest-growing sector of the telecommunications industry, providing service to over one billion customers worldwide. The exponential growth in number of subscribers started in early 1980s with commercial introduction of systems based on the cellular radio principles.

Wireless cellular communications has undergone three distinct phases of expansion, known as first, second, and third generations:

- First-generation (1G) systems, designed in the 1970s, and commercially introduced in early 1980s, are based on single-channel analog FM technology with channel spacings of 25 or 30 kHz. Such systems, which are still operating in parts of the world, are used almost exclusively for duplex voice transmissions.
- Second-generation (2G) systems were designed primarily in 1980s and commercially deployed in 1990s. They are all based on digital technology, making compact and power-efficient transceivers feasible. The primarily vehicular-mounted units of the first generation were, therefore, transformed into personal portable units of the second generation. Low-bit-rate data communication services are also introduced in 2G.
- Third-generation (3G) systems also operate on digital technology principles. Higher bandwidths allocated to 3G, combined with more sophisticated signal processing techniques, have greatly improved capacity and capabilities of wireless services. 3G is the gateway to personal multimedia, in which standards are capable of providing speech, data, video, and Internet services to wireless (mobile or fixed) units. International coordination for standardization of 3G has been performed by the ITU (International Telecommunications Union) in the framework of IMT-2000 (International Mobile Telecommunications) plan. Implementation of 3G systems started in 2002.

4.1.1. Major Second-Generation Standards. Major second-generation mobile radio standards that are also candidates for WLL are GSM, IS136, IS95 (high-power systems, originally developed for providing service to high-speed vehicular subscribers moving in outdoor large cell

environments), and DECT, PACS, and PHS (low-power microcellular systems, originally intended to provide coverage to low-speed outdoor pedestrians, and indoor users) [2]. Second-generation standards are reviewed by Black [12].

Detailed evaluations of DECT, PACS, and PHS standards have shown the suitability of all three for WLL applications [6]. Basic parameters of second generation standards are summarized below.

4.1.1.1. GSM (Global System for Mobile Communication). GSM, developed by CEPT (Conference Europenne des Postes et Telecommunications) in the 1980s to serve as a pan-European unified standard, evolved into a "model" digital mobile communications standard with the largest subscriber base in the world. GSM was standardized by ETSI (European Telecommunications Standard Institute). It operated in the FDD mode with 25 MHz bandwidth (935-960 MHz for downlink, and 890-915 MHz for uplink). Each band consists of 125 physical channels, each 200 kHz wide. Every physical channel operates in the TDMA mode with 8 user channels, each 0.577 ms. wide, forming 4.615-ms-long frames. Speech coding is RPE-LTP (regular pulse excited-long-term prediction) with 13 kbps (kilobits per second) for each subscriber. A combination of CRC and rate- $\frac{1}{2}$ convolution channel coding results in a gross bit rate per channel of 22.8 kbps. Modulation is Gaussian minimum shift keying (GMSK). GSM channel assignment is fixed (FCA).

The success of original GSM standard, and insufficiency of the original 900-MHz band resulted in introduction of DCS-1800 (Digital Communication System). This standard occupies a duplex pair of bands, each 75 MHz wide (1710–1785 MHz uplink, 1805–1880 MHz downlink). Each band, therefore, accommodates 375 channels with channel spacing of 200 kHz. All other parameters of DCS-1800 are identical to those of GSM. More details on GSM standard can be found in studies by Black [12] and Mehrotra [13].

4.1.1.2. North American TDMA Digital Cellular (IS136). The TDMA-based IS136 standard was developed by TR45.3, a subcommittee of the EIA/TIA (Electronic Industry Association/Telecommunications Industry Association) in the United States. IS136 is compatible with the analog, first-generation system AMPS (advance mobile phone service). It operates in two frequency bands (869-894 MHz uplink, 824-849 MHz downlink). There are 832 physical channels with a channel spacing of 30 kHz (the same as analog FM channels used in AMPS). Channel assignment is FCA. Each channel accommodates three users in TDMA mode with a channel bit rate of 48.6 kbps. TDMA frame length is 40 ms, speech coding is VSELP (vector sum excited linear predictive), channel coding is rate- $\frac{1}{2}$ convolution, and modulation is $\pi/4$ -DQPSK (differential quadrature phase shift keying). The standard provides flexibility to accommodate six users in the same 30-kHz band. Black has reported the details of this standard [12].

4.1.1.3. North American CDMA Digital Cellular Cdma-One (1895). This is the first CDMA digital cellular standard in the world, developed by Qualcomm Inc. in the United States, and standardized by Subcommittee TR45.5 of the EIA/TIA in 1993. The standard uses the same 800-MHz band as analog AMPS and digital TDMA IS136 standards. Each duplex band of 25 MHz is, however, divided into 20 channels, each 1.25 MHz wide. Each channel provides service to a number of users, which are each assigned a distinct code. The technique is based on direct-sequence spread spectrum in which the 9.6-kbps user data are converted to 1.2288 Mcps (million chips per second), occupying one 1.25-MHz physical channel. Frame length is 20 ms, speech coding is QCELP (Qualcomm code excited linear predictive), channel coding is rate- $\frac{1}{2}/\frac{1}{2}$ convolution in the downlink/uplink paths, and modulation is OQPSK (offset QPSK). Channel assignment of IS95 is DCA. More details of the standard are reported by Black [12].

4.1.1.4. DECT (Digital Enhanced Cordless Telecommunications). DECT is a European standard also developed by ETSI. It is designed to operate in the 1880-1900-MHz frequency band, with flexibility to use other close bands. It is based on the TDMA-TDD principle. The number of carriers is 10, and carrier separation is 1726 kHz. The transmission rate is 1152 kbps, and the number of TDMA channels for each carrier is 12. Therefore, the total number of voice channels is 120 (10 carriers \times 12 time slots per carrier). Speech coding is 32 kbps ADPCM (adaptive differential pulse code modulation), and modulation method is GFSK (Gaussian frequency shift keying). Channel assignment is dynamic. Maximum transmission power of the base and portable is 250 mW, where dynamic power control reduces it down to 60 mW. This, however, is peak power used during the transmission of a time slot. Average power is ≤ 10 mW, resulting in long battery usage before recharge. Normal cell radius in DECT is several hundred meters. For each voice connection two time slots are used for two-way transmission. In the other 22 time slots the portable unit scans and evaluates other channels for handover to a better channel when available. More information about DECT can be found in the article by Yu et al. [14].

4.1.1.5. PACS (Personal Access Communication System). PACS has been developed in the United States and was standardized by the JTC (Joint Technical Committee) in 1994. It operates in two wide duplex bands 1850–1910 MHz (uplink) and 1930–1990 MHz (downlink). These bands were allocated by the FCC (Federal Communications Commission) in three paired 5-MHz and three paired 15-MHz bands for licensed wideband PCS applications. Also a 10-MHz band (1920–1930 MHz) has been allocated for unlicensed TDD operation. The air interface of PACS allows FDD operation in the licensed band and TDD operation in the unlicensed band [15].

The PACS standard is based on FDD-TDMA (frequency-division duplex) with 200 channels (carrier separation of 300 kHz). Modulation and speech coding are $\pi/4$ -QPSK (quadrature phase shift keying) and 32 kbps ADPCM (adaptive pulse code modulation), respectively. Bit rate per channel is 384 kbps.

	Hig	h-Power Macrocel	lular	Lo	ow-Power Microcellula	r
System	GSM	IS136	IS95	DECT	PACS	PHS
Frequency band (MHz)	935 - 960 890 - 915	869 - 894 824 - 849	869 - 894 824 - 849	1880-1900	1850 - 1910 1930 - 1990	1895-1918
Standardization body	ETSI	EIA/TIA TR45.3	EIA/TIA TR45.5	ETSI	JTC	ARIB
Duplex method	FDD	FDD	FDD	TDD	FDD	TDD
Access method	FDMA/TDMA	FDMA/TDMA	FDMA/CDMA	FDMA/TDMA	FDMA/TDMA	FDMA/TDMA
Number of carriers	124	832	20	10	200	77
Carrier separation (kHz)	200	30	1250	1728	300	300
Modulation	GMSK	$\pi/4$ -DQPSK	QPSK	GFSK	$\pi/4$ -QPSK	$\pi/4$ -QPSK
Rate per channel (kbps)	270.83	48.6	1228.8	1152	384	384
Frame time (ms)	4.615	40	20	5 + 5	2.5	2.5 + 2.5
Slots per frame	8/16	3/6	1	12 + 12	8	4 + 4
Speech coding type and rate (kbps)	RPE-LTP, 13	VSELP, 7.95	QCELP, 9.6	ADPCM, 32	ADPCM, 32	ADPCM, 32
Channel coding	Rate- $\frac{1}{2}$ convolution	$\begin{array}{c} \text{Rate-}\frac{1}{2} \\ \text{convolution} \end{array}$	$\frac{1}{2}$ forward $\frac{1}{3}$ reverse	CRC	CRC	CRC
Channel assignment	FCA	FCA	DCA	DCA	QSAFA	DCA
Modulation efficiency (bps/Hz)	1.35	1.62	0.98	0.67	1.28	1.28
Handoff strategy	Mobile-assisted	Mobile-assisted	Mobile-assisted	Mobile-controlled	Mobile-controlled	Mobile-assisted

Table 1. Basic Parameters of the Second-Generation Wireless Standards

Channel assignment is quasistatic autonomous frequency assignment (QSAFA/DCA) [15]. The standard is designed for low mobility applications. However, operation at high speed (several tens of kilometers per hour) is also possible. Maximum transmission power of the portable unit is 200 mW, and average power is 25 mW. More details about PACS have been reported [14,15].

4.1.1.6. *PHS* (*Personal Handy-Phone System*). PHS is the Japanese-developed standard operating in the 1895–1918-MHz band. PHS was envisioned as an efficient low-cost cordless and portable phone system. In late 1993 RCR (Research and development Center for Radio systems), currently known as ARIB (Association of Radio Industries and Businesses), approved the RCR STD-28 standard. The interface for connection to the network was subsequently completed by the TTC (Telecommunication Technology Committee), and trial systems started operation. The first commercial system was implemented in mid-1995. Since then, PHS has experienced explosive growth in Japan, reaching a market size of over 8 million in 1999.

PHS is based on the TDMA-TDD principle with 77 channels (carrier separation of 300 kHz). Bit rate is also 384 kbps, and modulation is $\pi/4$ -QPSK. Speech coding is 32 kbps ADPCM and channel assignment is dynamic. Each physical channel can be used as four traffic

channels in the TDMA mode. Details of PHS are reported in Ref. 16.

Table 1 summarizes parameters of the six major highpower and low-power second-generation digital cellular standards.

4.1.2. Major Third-Generation Standards. The unparalleled success and exponential growth in first-generation mobile communication systems in the early 1980s necessitated initiation of collective efforts in developing standards that could be used internationally to realize the slogan of PCS (personal communication services), wireless access of "any kind, to any one, and at any where." Such efforts were initiated at the ITU in 1985 in the framework of FPLMTS (future public land mobile telephone systems), which was later renamed IMT-2000. The goals set at IMT-2000 was to provide flexible and spectrum efficient voice and data services to wireless users. The minimum bit rate requirements for outdoor high-speed macrocellular, outdoor pedestrian microcellular, and indoor picocellular environments are 144 kbps, 384 kbps, and 2 Mbps, respectively.

To satisfy the needs of a large international subscriber base requires two-way transmission of low-to-high bit rate data and video in addition to conventional voice telephony. The ITU allocated 230 MHz in the 2 GHz band at WARC'92 (World Administrative Radio Conference). The frequency bands are 1885–2025 MHz and 2110–2200 MHz. To provide global coverage and roaming capabilities, both terrestrial and satellite links were considered.

Many proposals were submitted to the ITU, and a number of standards capable of fulfilling the IMT-2000 vision emerged in the late 1990s. The three major standards, two of them based on wideband CDMA, and one on wideband TDMA, are described in this section. The degree of suitability of each standard for WLL applications is yet to be determined. However, the anticipated largescale deployment of equipment based on these standards, which results in introduction of economically attractive systems, coupled with the capability to provide new higher bit rate services, make these standards suitable candidates for future WLL systems.

Major third-generation standards have been described [17,18].

4.1.2.1. European-Based WCDMA (Wideband CDMA).

This standard, which is also supported by the Japanese wireless industry, has been developed by ETSI in the framework of the European UMTS (Universal Mobile Telecommunication Systems) project. UMTS is the European version of IMT-2000. WCDMA operates in the paired band of the IMT-2000 spectrum based on FDD. It uses 5-, 10-, 15-, and 20-MHz-wide channels, and operates at chip rates of 1.024, 4.096, 8.192, and 16.384 Mcps. Frame length is 10 ms. Provisions are made for multirate and packet data services. The standard is backward-compatible with GSM. Yang has described the basic principles and applications of CDMA [19]. Detailed descriptions of WCDMA are provided by other authors [20,21].

4.1.2.2. North American WCDMA Standard Cdma2000. The cdma2000 standard was finalized by the Subcommittee TR45.5 of the TIA Engineering Committee TR45 in the United States in March 1998. It is backward-compatible with the cdmaOne (IS95) standard. Provisions for packet data transmission are provided, channel bandwidths are $N \times 1.25$ MHz (N = 1, 4, 8, 12, 16), and chip rates are 1.2288, 3.6864, 7.3728, 11.0593, and 14.7456 Mcps. Frame length is 20 ms.

4.1.2.3. North American UWC-136 (Universal Wireless Communications). The UWC-136 standard, prepared by Subcommittee TR45.3 of the TIA, is a family of technologies based on TDMA. It consists of (1) the IS136 standard with 30-kHz channels for speech and data below 28.8 kbps; (2) IS136+, which again uses 30-kHz channels for speech and data (bit rates, however, are increased to 64 kbps applying *M*-ary modulation; (3) IS136 HS (high speed) for outdoor vehicular applications using 200 kHz wide channels (data rates of \leq 384 kbps are possible; duplex policy is FDD); and (4) IS136 HS indoor, in which the 1.6-MHz-wide channels make data rates up to 2 Mbps feasible. FDD and TDD methods are both used. IS136 HS is also compatible with GSM, using the same frame length of 4.615 ms.

The basic parameters of major 3G standards are summarized in Table 2. Further details can be found in the literature [17,18].

4.2. Proprietary Radio Interface Standards

These standards are developed by private organizations to replace the existing wired loops. Such standards cover a wide range of carrier frequencies, radio interface technology, transmission rates, range, performance, and types of service. Major standards cited by the ITU are Nortel Proximity I-Series, SR Telecom's SR 500, and TRT/Lucent Technologies IRT [2]. Other major systems cited by Webb [3] are Innowave Multigain, Airspan, Lucent AirLoop, Interdigital Broadband CDMA, and Granger CD2000. An overview of these standards and

|--|

				UWC-136	
System	WCDMA	Cdma2000	IS136+	IS136 HS Outdoor/Vehicular	IS136HS Indoor
Backward compatibility	GSM	CdmaOne (IS95)		IS136	
Standardization body	ETSI	EIA/TIA TR45.5		EIA/TIA TR45.3	
Access method	WCDMA	WCDMA		TDMA	
Carrier separation	5, 10, 20 MHz	$1.25, 5, 10, 15, 20 \mathrm{~MHz}$	$30 \mathrm{kHz}$	200 kHz	$1.6 \; \mathrm{MHz}$
Maximum user bit rate per code or channel	480, 960, 1920 Kbps	1.0368 Mbps	64 Kbps	384 Kbps	2 Mbps
Maximum user bit rate (multicode)	2 Mbps	2 Mbps		N/A	
Chip rate (Mcps)	1.024, 4.096, 8.192, 16.384	1.2288, 3.6864, 7.3728, 11.0593 direct spread $n \times 1.2288$ (n = 1,3,6,9,12) for multicarrier		N/A	
Frame length (ms)	10	20	40	4.615	4.615

details of their basic parameters have been presented by the ITU [2] and Webb [3].

5. WLL SPECTRUM AND CAPACITY

Large-scale deployment of WLL depends on availability and efficient use of the radio spectrum. Wireless access systems, fixed and mobile, operate in the wide frequency range of 400 MHz-40 GHz. A number of analog FM systems based on European standards operate at 400 MHz. IS136 mobile cellular uses the 800-MHz band, while GSM operates at 900 MHz. Point-to-multipoint radio in the loop has been designed at 1.4 GHz. The 1.8-1.9-GHz bands are used for GSM, DECT, and PHS standards. A wide band at 2 GHz has been allocated by the ITU for global implementation of IMT-2000 standards. Multipoint distribution systems (MDSs) and point-to-multipoint radio also operate at 2.4, 2.5, 2.6, and 10.5 GHz. Finally, 28- and 40-GHz bands are used for local multipoint communications and distribution systems (LMCS/LMDS) [2]. Most of these frequencies, especially lower bands, have been developed for mobile users; however, they are either being used, or have the potential of being used for WLL-type applications.

Capacity is the most critical issue in wireless personal communications, fixed or mobile. Cellular radio architecture provides high capacities because of its spectrum efficiencies [22]. It has been shown by an example that even the relatively low capacity cellular analog FM standards are capable of providing millions of fixed WLL-type telephone lines in a metropolitan area if a reasonably large frequency band is allocated [5]. Capacity of digital cellular systems is higher than analog because (1) speech coding reduces bandwidth per subscriber, and (2) channel coding results in smaller signal-to-interference requirements, which in turn decreases minimum reuse distance, and further increases capacity.

Performance and capacity of DECT, PACS, and PHS standards for WLL applications have been reported [6]. Detailed qualitative and quantitative evaluations indicate that all three standards provide satisfactory performance for WLL applications. For low-traffic environments, PACS which can employ larger cells performs better than the other two standards. In suburban areas where in addition to coverage capabilities capacity is an issue, DECT has better performance. For high-traffic-density urban areas with great capacity requirements, the three standards all have good performance [6].

TDMA systems have higher capacity than do FDMA systems. In CDMA more interference can be tolerated, which boosts capacity even higher. The CDMA-versus-TDMA capacity of cellular mobile radio systems was a subject of intense debate in the 1990s. Real capacity evaluations are difficult due to a large number of parameters and assumptions. It has been suggested [3] that capacity of CDMA for WLL applications is 1.4–2.5 times that of TDMA. Cell sectorization results in even further CDMA capacity improvements.

Capacities of cellular mobile and fixed WLL have been compared with reference to detailed calculations [10,11]. In one study, the capacities of the IS136 (TDMA), GSM (TDMA), and IS95 (CDMA) standards were evaluated in detail and compared for both fixed and mobile access [10]. CDMA WLL capacity (measured in erlangs per cell per MHz) was found to be about 2.5 times that of IS136, and about 5.4 times that of GSM. Detailed analysis [11] has shown that capacity of WLL is not necessarily higher than that of mobile if FDMA or TDMA is used. It is higher if CDMA is applied. For WLL applications alone, CDMA always provides higher capacity than does TDMA [11].

Many advantages of CDMA have made it the major access method for the emerging third-generation wireless mobile standards. CDMA is also expected to be access of the choice for WLL applications in years to come.

6. ECONOMICS OF WLL

Successful deployment and operation of any communication system depends on its economic viability. In a typical system the operator makes an initial capital investment to build the initial infrastructure for a system that will grow and provide service to a large number of subscribers in the future. Interest should be paid on the capital sum until the network becomes profitable. Revenues increase as customer base increases. There are also ongoing maintenance, upgrade, and marketing costs associated with operation of the system.

A typical communication system consists of a core network of switching and transmission equipment for backhaul connections, and the access part, which connects the subscriber to the network. The network cost is approximately the same for wired and wireless loops. The per subscriber cost of the core network, estimated at 30-60 for a large network of half a million users [3], is insignificant as compared to the cost of access. Detailed cost evaluations are complex for both wired and wireless access (especially wired) because of the large number of interrelated parameters. It is also changing with time as a result of inflation and changes in technology. However, one can compare the two by elaborating on the involved parameters of each, and by looking at their trends.

6.1. Cost of Wired Loops

The copper-based wired loops require lumps of investments to lay down main cables to support a future large subscriber base. The cost of unused capacity and the interest payments on the initial large capital makes competition stiff for small-size new operators. The cost per line of the access varies greatly, depending on the number of customers, length of loop, and type of terrain. As an example, analysis of cable costs for 30 rural area sites in the United States in 1990 showed a per line cost as low as \$834 and as high as \$50,000. The average per line cost for all 30 sites covering 8282 subscribers was \$3102 [1]. Looking at the trends, however, it can be shown that cost per access line increases almost linearly as a function of length of access line, up to a point where it takes a jump as a result of required loading coils. From that point it increases almost linearly, although with a larger slope due to the required shift to larger gauge cable. Increasing the distance results in another jump for insertion of additional loading coils,

and a subsequent exponential increase afterward due to even larger gauge cable, and due to the fact that very long loops normally involve higher costs because of difficult terrain [1].

The per line cost of wired loops in a flat area normally decreases as penetration (subscriber density) increases. Simple calculations by Webb [3] indicate that in an environment where houses are located adjacent to each other, cost per subscriber increases from \$450 to \$1650 when penetration decreases from 100% to 20%.

6.2. Cost of Wireless Loops

In wireless local loops, investments are made in small increments, matching growth in traffic (Fig. 2). The large idle initial investment can, therefore, be avoided. Subsequent interest payment on investment is also considerably smaller. Installation costs are divided typically 20% on infrastructure (initial investment for base station equipment) and 80% on subscriber (paid when the customer receives service, which is immediately followed by generated revenue). These factors make WLL particularly attractive for low-capital new-entrant operators.

The cost of wireless access also depends on a number of factors, including type of access and subscriber density. Per subscriber costs are lowest for low-power microcellular systems operating in dense outdoor urban areas and in indoor environments. The cost is highest for megacellular satellite systems providing coverage to low-density rural areas. The most important aspect of WLL cost is that it is distance-insensitive. If a subscriber is within the coverage area of a base station, its distance to the base does not affect cost.

The cost of WLL can be divided into "shared" and "dedicated." Shared costs are those allocated to a number of subscribers, such as base station equipment. Dedicated costs are per subscriber costs such as terminal transceiver and antenna. Calhoun [1] subdivided WLL costs into the following five major categories:

- 1. *Common equipment costs*, such as power supplies, base station antennas, and central processor for system control. This type of cost does not vary with the size of the system.
- 2. *Traffic-sensitive equipment costs* costs mainly for RF channel hardware (base station transceiver or radio units), the cost of which depends on the number of subscribers and average traffic statistics.
- 3. *Per subscriber equipment costs* costs related to equipment purchased one unit per subscriber. This includes subscriber transceiver, and line card to interface the central office.
- 4. Ancillary materials and labor—includes costs for the pole, antenna, and power supply, all installed in customer premises, and dedicated to a single customer, or shared by a number of customers at the same location.
- 5. Overhead charges costs charged as a standard percentage of the total value of a project.

This cost structure is depicted in Fig. 3. A simple but effective formula to estimate cost of wireless loops is also provided by Calhoun [1]. The total project $\cos t R$ is given by R = AX + BY + C, where X is the number of subscribers and Y is the number of radio channels. A, B, and C are equipment cost of the per subscriber equipment (item 3 above), equipment cost of per RF-channel equipment (item 2), and cost of the common equipment (item 1), respectively. The relationship between X and Y is not fixed; it depends on the user calling habits (average holding times) and on the grade of service (blocking probability).



Figure 3. Major components of WLL for cost categorization. (Source: Ref. 1.)

Using this approach, one can deduce that total per subscriber cost is high for small X; it decreases rapidly as X increases, and reaches a lower limit (saturation point) where cost of subscriber station plus traffic-sensitive base station cost are dominant cost factors.

6.3. Comparison Between Wired and Wireless Costs

The costs of wired and wireless loops are both decreasing functions of subscriber density (number of subscribers per square kilometer). Comparison of wireless and wireline systems in terms of capital (installation) cost is provided in Fig. 4. Cost per subscriber of WLL in the 1990s was lower than in the 1980s because of mass production of cellular equipment and the "negative inflation" phenomena associated with electronics. Figure 4 shows a breakpoint around 100 subscribers per square kilometer. As time passes, this breakpoint is expected to shift even further to the right (i.e., WLL becomes economically superior for even larger subscriber densities). The reason is that wired loop economy depends on the cost of raw material (copper) and labor, both of which increase with time. WLL economy on the other hand, is governed by the "negative inflation" phenomena associated with electronics.

A case study comparing economy of two access methods is provided by Webb [3]. Three flat areas representing a range of housing densities (high-density, medium-density, and low-density) were considered. It has been shown that in each case when penetration (fraction of houses subscribing to the service) decreases from 25% to 10%, cost per subscriber of cable access increases almost linearly with a modest slope. When penetration decreases below 10%, slope of the cost curve rises sharply. The cost of WLL, however, remains almost the same for penetrations of 5-25%. This comparative study also indicates that per subscriber cost of wired loops is much more sensitive to housing density, as compared to the cost of wireless loops.

Another study shows that a combination of wired and wireless loops provides best economic solutions for



Figure 4. Cost versus subscriber density for wireline and wireless access systems. (*Source*: Ref. 2.)

a number of cases [1]. In this study relative cost is plotted as a function of percentage of wireless loops, ranging from 0 to 100%. There is an optimum percentage point where the cost is lowest. The optimum point, however, is very sensitive to the individual case, and to the governing parameters and assumptions. It is important to note that capital (installation) cost is only one aspect of economic feasibility. A valid comparison of the two access methods should be based on "annual lifetime cost," which contains capital cost, operating cost, and replacement cost. Such analysis points further at the superiority of WLL [2]. It is estimated that up to 80% of a telephone company's total maintenance cost is allocated to the local loop. WLL represents great savings in maintenance costs since the expensive operations associated with digging the ground and replacing wires are totally eliminated. Operating expenses can be reduced by as much as 25% per subscriber per year in WLL [2]. Considering lifetime cost per subscriber instead of installation costs alone makes WLL superior to wireline-based networks for subscriber densities below approximately 200-400 subscribers per square kilometer. The exact position of the crossover point depends on specific assumptions, distribution of subscribers, and traffic levels.

A major conclusion is that "Wireless access based systems are cost-effective for the provision of telephony services to typical residential subscribers, particularly in rural and suburban areas, or for new entrants in competitive urban markets" [2]. Economic feasibility, combined with other advantages described before, makes WLL access the access of choice in most cases. Large-scale deployment, however, is subject to availability of radio spectrum.

7. BROADBAND APPLICATIONS OF WLL

For over 100 years telecommunication systems have been dominated by "telephony," that is, two-way transmission of voice. More recently, however, delivery of broadband services such as high-speed Internet and video to the home and office has become increasingly important. Fixed wireline operators are considering digital subscriber techniques such as ADSL (asymmetric digital subscriber line) to increase capacity of copper lines [23]. Very-high-bit-rate services to subscribers can also be delivered by fiberopticbased techniques such as FTTC (fiber to the curb) and FTTH (fiber to the home). These techniques, however, share many disadvantages with the copper-based local loops.

Subject to allocation of sufficient radio spectrum, WLL is also capable of providing broadband services. It should be noted that "broadband" is not a well-defined term. In voice-oriented mobile telephony, transmission rates above 100 Kbps are considered as "high bit rates." For cable operators, broadband is 8 Mbps or higher.

Technical and economical aspects of broadband WLL, along with a forecast of future are provided by Webb [3]. Third-generation mobile wireless standards provide a peak data rate of 2 Mbps in indoor picocellular environments at the 2-GHz band. This rate provides limited

2958 WIRELESS LOCAL LOOP STANDARDS AND SYSTEMS

multimedia capabilities. The emerging new broadband services, however, require higher data rates. As an example, in 1997 the ETSI established the BRAN (Broadband Radio Access Network) project in Europe [24]. The purpose of the project was to utilize broadband LAN (local-area network) technology and broadband fixed radio access to provide mechanisms for delivery of multimedia services to subscribers. In the framework of the BRAN project HiperAccess was suggested for WLL systems. The BRAN project came to the conclusion that a peak data rate of 25 Mbps is sufficient for most broadband user-oriented applications (1.5-6 Mbps for video applications, 2 Mbps for Web browsing, 10 Mbps for corporate access, and up to 25 Mbps for LAN connections). Such high bit rates, however, require transmission at much higher frequencies (say, above 10 GHz), for which technology is evolving. Simple calculations [3] have shown that for economic viability of broadband WLL services, assignment of radio spectrum at least 10 times the bandwidth offered to a user is required. A major conclusion is that "for typical assignments in the frequency bands of 10 GHz and above, bandwidths of 10 MHz per subscriber, using WLL would seem readily achievable" [3]. A number of standards for broadband WLL are being developed, and a number of broadband proprietary WLL products are being introduced to the market [3]. It is safe to expect widespread deployment of broadband WLL-based systems and services in the near future.

8. CONCLUSIONS

In this tutorial presentation the basic principles and applications of fixed wireless access, or WLL, were reviewed. WLL technology, standards, spectrum efficiency, capacity, and economics were described.

A major conclusion is that WLL is an efficient and economically feasible alternative for wired local loops. Major second-generation mobile radio standards, as well as proprietary radio interface standards, have been deployed in WLL applications. Third-generation mobile radio standards and emerging broadband WLL standards, mostly based on wideband CDMA, are potential candidates for providing voice and data services to subscribers in highdensity urban areas, as well as sparsely populated rural areas.

BIOGRAPHY

Homayoun Hashemi received the B.S.E.E. degree from the University of Texas at Austin in 1972, and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Sciences, and the M.A. degree in Statistics, all from the University of California at Berkeley, in 1974, 1977, and 1977, respectively. He joined Bell Telephone Laboratories, Holmdel, New Jersey in 1977, where he was involved in system design for high-capacity mobile telephone systems. Since 1979 he has been a faculty member at Sharif University of Technology in Teheran, Iran, where he is currently a full Professor of Electrical Engineering. Dr. Hashemi has done research on different aspects of wireless communications, with emphasis on propagation modeling. His channel simulator package SURP has been used internationally in the design of digital cellular radio communication systems. He spent one year at NovAtel Communications Ltd. in Calgary, Canada in 1990, the summers of 1992 and 1994 at the Electrical Engineering Department of the University of Ottawa, and the summer of 1993 at TRLabs in Calgary. During these visits he defined and supervised three major projects; in each of these projects the largest propagation database of its kind in the world, even to this date, was set up and analyzed. He received the "Best Paper Award" at the IEEE, VTC'99 Conference, and the IEEE Vehicular Technology Society's Neal Shepherd Best Propagation Paper Awards in 2000 and 2001.

BIBLIOGRAPHY

- 1. G. Calhoun, Wireless Access and the Local Telephone Network, Artech House, Norwood, MA, 1992.
- International Telecommunications Union, Handbook on Land Mobile (Including Wireless Access), Vol. I: Wireless Access Local Loop, Radio Communication Bureau, 1996.
- 3. W. Webb, Introduction to Wireless Local Loop, 2nd ed., Broadband and Narrowband Systems, Artech House, 2000.
- 4. ITU, Draft Revision of Recommendation ITU-R F.757, Basic System Requirements and Performance Objectives for Cellular Type Mobile Systems Used as Fixed Systems (Fixed Wireless Local Loop Applications of Cellular Type Mobile Technologies), Document 9B/73, Jan. 1997.
- H. Hashemi, K. Anvari, and M. Tabiani, Application of cellular radio to telecommunication expansion in developing countries, *Proc. IEEE GLOBECOM'92 Conf.*, Orlando, FL, Dec. 6-9, 1992, pp. 984-988.
- 6. O. Momtahan and H. Hashemi, A comparative evaluation of DECT, PACS, and PHS standards for wireless local loop applications, *IEEE Commun. Mag.* **39**(5): 156–163 (May 2001).
- 7. V. H. MacDonald, The cellular concept, *Bell Syst. Tech. J.* **58**(1): 15–41 (Jan. 1979).
- 8. W. C. Y. Lee, Mobile Cellular Telecommunications Systems, Analog and Digital, McGraw-Hill, New York, 1995.
- 9. T. S. Rappaport, Wireless Communications, Principle and Practice, Prentice-Hall, 1996.
- V. K. Garg and E. L. Sneed, Digital wireless local loop system, IEEE Commun. Mag. 34: 112–115 (Oct. 1996).
- W. C. Y. Lee, Spectrum and technology of a wireless local loop system, *IEEE Pers. Commun. Mag.* 5: 49–54 (Feb. 1998).
- U. Black, Second Generation Mobile & Wireless Technologies, Prentice-Hall, 1998.
- 13. A. Mehrotra, GSM System Engineering, Artech House, 1997.
- Ch. C. Yu et al., Low-tier wireless local loop systems Part I: Introduction, *IEEE Commun. Mag.* 35: 84-92 (March 1997).
- A. R. Noerpel, Y. B. Lin, and H. Sherry, PACS: Personal communications system—a tutorial, *IEEE Pers. Commun. Mag.* 3: 32-43 (June 1996).
- 16. S. Sampei, Application of Digital Wireless Technologies to Global Wireless Communications, Prentice-Hall, 1997.
- 17. P. Stavroulakis, Third Generation Mobile Telecommunication System: UMTS & IMT-2000, Springer-Verlag, 2000.

- 18. R. Prasad, *Third Generation Mobile Communication Systems*, Artech House, 2000.
- S. C. Yang, CDMA RF System Engineering, Artech House, 1998.
- 20. T. Ojanpera and R. Prasad, Wideband CDMA for Third Generation Mobile Communications, Artech House, 1998.
- H. Holma and A. Toskala, eds., WCDMA for UMTS, Radio Access for Third Generation Mobile Communications, Wiley, New York, 2000.
- W. C. Y. Lee, Spectrum efficiency in cellular, *IEEE Trans. Vehic. Technol.* 38(2): 69-75 (May 1989).
- P. Kyees et al., ADSL: A new twisted pair access to the information highway, *IEEE Commun. Mag.* 33: 52-60 (April 1995).
- J. Haine, HiperAccess: An access system for the information age, *IEE Electron. Commun. Eng. J.* 10(5): 229-235 (Oct. 1998).

WIRELESS LOCATION

ALI H. SAYED NABIL R. YOUSEF Adaptive Systems Laboratory University of California Los Angeles, California

1. **DEFINITION**

Wireless location refers to obtaining the position information of a mobile subscriber in a cellular environment. Such position information is usually given in terms of geographic coordinates of the mobile subscriber with respect to a reference point. Wireless location is also commonly termed *mobile positioning*, *radiolocation*, and *geolocation*.

2. APPLICATIONS

Wireless location is an important public safety feature of future cellular systems since it can add a number of important services to the capabilities of such systems. Among these services and applications of wireless location are [e.g., 1-11]:

1. *E-911*. A high percentage of emergency 911 (E-911) calls nowadays come from mobile phones [1,2]. However, these wireless E-911 calls do not get the same quality of emergency assistance that fixed-network E-911 calls enjoy. This is due to the unknown location of the wireless E-911 caller. To face this problem, the Federal Communications Commission (FCC) issued an order on July 12, 1996 [1], which required all wireless service providers to report accurate mobile station (MS) location to the E-911 operator at the public safety answering point (PSAP). According to the FCC order, it is mandated that within 5 years from the effective date of the order, October 1, 1996, wireless service providers must convey to the PSAP the location of the MS within 100 m of its actual location for at least 67% of all wireless E-911 calls.¹ It is also expected that the FCC will further tighten the required location accuracy level in the near future [3]. This FCC mandate has motivated research efforts toward developing accurate wireless location algorithms and in fact has led to significant enhancements to the wireless location technology [e.g., 4-11].

2. Location-Sensitive Billing. Using accurate location information of wireless users, wireless service providers can offer variable-rate call plans that are based on the caller location. For example, the cell-phone call rate might vary according to whether the call was made at home, in the office, or on the road. This will enable wireless service providers to offer competitive rate packages to those of wire-line phone companies.

3. Fraud Protection. Cellular phone fraud has attained a notorious level, which serves to increase the usage and operation costs of cellular networks. This cost increase is directly passed to the consumer in the form of higher service rates. Furthermore, cellular fraud weakens the consumer confidence in wireless services. Wireless location technology can be effective in combating cellular fraud since it can enable pinpointing perpetrators.

4. Person/Asset Tracking. Wireless location technology can provide advanced public safety applications including locating and retrieving lost children, Alzheimer patients, or even pets. It could also be used to track valuable assets such as vehicles or laptops that might be lost or stolen. Furthermore, wireless location systems could be used to monitor and record the location of dangerous criminals.

5. *Fleet Management*. Many fleet operators, such as police force, emergency vehicles, and other services including shuttle and taxicab companies, can make use of the wireless location technology to track and operate their vehicles in an efficient way in order to minimize response times.

6. *Intelligent Transportation Systems*. A large number of drivers on road or highways carry cellular phones while driving. The wireless location technology can serve to track these phones, thus transforming them into sources of real-time traffic information that can be used to enhance transportation safety.

7. Cellular System Design and Management. Using information gathered from wireless location systems, cellular network planners could improve the cell planning of the wireless network based on call/location statistics. Improved channel allocation could be based on the location of active users [9,10].

8. *Mobile Yellow Pages*. According to the available location information, a mobile user could obtain road information of the nearest resource that the user might need such as a gas station or a hospital. Thus, a cellular phone will act as smart handy mobile yellow pages on demand. Cellular users could obtain real-time traffic information according to their locations.

 $^1\,{\rm The}\,$ original FCC requirement was $125\,{\rm m}\,$ and was then tightened to 100 m.

3. WIRELESS LOCATION TECHNOLOGIES

Wireless location technologies fall into two main categories: mobile-based and network-based techniques. In mobile-based location systems, the mobile station determines its own location by measuring signal parameters of an external system, which can be the signals of cellular base stations or satellite signals of the Global Positioning System (GPS). On the other hand, network-based location systems determine the position of the mobile station by measuring its signal parameters when received at the network cellular base stations. Thus, in the later type of wireless location systems, the mobile station plays a minimal or no role in the location process.

3.1. Mobile-Based Wireless Location

3.1.1. GPS Mobile-Based Location Systems. In GPS-based location systems, the MS receives and measures the signal parameters of at least four different satellites of a currently existing network of 24 satellites that circle the globe at an altitude of 20,000 km and which constitute the Global Positioning System. Each GPS satellite transmits a binary code, which greatly resembles a code-division multiple-access (CDMA) code. This code is multiplied by a 50-Hz unknown binary signal to form the transmitted satellite signal. Each GPS satellite periodically transmits its location and the corresponding timestamp, which it obtains from a highly accurate clock that each satellite carries. The satellite signal parameter, which the MS measures for each satellite, is the time the satellite signal takes until it reaches the MS. Cellular handsets usually carry a less accurate clock than the satellite clock. To avoid any errors resulting from this clock inaccuracy, the MS timestamp is often added to the set of unknowns that need to be calculated, thus making the number of unknowns equal to four (three MS position coordinates plus timestamp). This is why four satellite signal parameters have to be measured by the MS. Further information on the GPS systems is available in the literature [12,13].

After measuring the satellite signal parameters, the MS can proceed in one of two manners. The first is to calculate its own position and then broadcast this position to the cellular network. Processing the measured signal parameter to obtain a position estimate is known as *data fusion*. In the other scenario, the MS broadcasts the unprocessed satellite signal parameters to another node (or server) in which the data fusion process is performed to obtain an estimate of the MS position. The later systems are known as *server-aided* GPS systems, while the first are known as "pure" GPS systems [14,15].

A general scheme for server-aided GPS systems is shown in Fig. 1. The server-aided GPS approach is successful in a microcell cellular environment, where the diameter of cellular cells is relatively small (a few hundred meters to a few kilometers). This environment is common in urban areas. On the other hand, in macrocell environments, which are common in suburban or rural areas, base stations, and thus servers, are widely spread out. This increases the average distance between the MS and the server leading to ineffective



Figure 1. Server-aided GPS location.

correction information. This is why, in many mobile-based GPS location system designs, handsets have to support both server-aided GPS and pure GPS location modes of operation [e.g., 14].

GPS-based mobile location systems have the following advantages. GPS receivers usually have a relatively high degree of accuracy, which can reach less than 10 m with differential GPS server-aided systems [16]. Moreover, the GPS satellite signals are available all over the globe, thus providing global location information. Finally, GPS technology has been studied and enhanced for a relatively long time and for various applications, and is a rather mature technology. Despite these advantages, wireless service providers may be unwilling to embrace GPS fully as the principal location technology due to the following disadvantages of GPS-based location systems:

- 1. Embedding a GPS receiver in the mobile handset directly leads to increased cost, size, and battery consumption of the mobile handset.
- 2. The need to replace hundreds of millions of handsets that are already in the market with new GPSaided handsets. This will directly impact the rates that the wireless carriers offer their users and can cause considerable inconvenience to both users and carriers during the replacement period.
- 3. The degraded accuracy of GPS measurements in urban environments, when one or more satellites are obscured by buildings, or when the mobile antenna is located inside a vehicle.
- 4. The need for handsets to support both serveraided and pure GPS modes of operation, which increases the average cost, complexity, and power consumption of the mobile handset. Furthermore, the power consumption of the handset can increase dramatically when used in the pure GPS mode. Moreover, the need to deploy GPS aiding servers in wireless base stations adds up to the total cost of GPS-aided location systems.
- 5. GPS-based location systems face a political issue raised by the fact that the GPS satellite network is controlled by the U.S. government, which reserves the right to shut GPS signals off to any given region worldwide. This might make some wireless service providers outside the United States unwilling to rely solely on this technology.

3.1.2. Cellular Mobile-Based Location Systems. Cellular mobile-based wireless location technology is similar to

GPS based location technology, in the sense that the MS uses external signals to determine its own location. However, in this type of location systems, the MS relies on wireless signals originating from cellular base stations. These signals could be actual traffic cellular signals or special-purpose probing signals, which are specifically broadcast for location purposes. Although this approach, which is also known as forward-link wireless location, avoids the need for GPS technology, it has the same disadvantages that GPS location systems have, which is the need to modify existing handsets, and may even have increased handset power consumption over that of the GPS solution. In addition, this solution leads to lower location accuracy than that of the GPS solution. This makes cellular mobile-based location systems less favorable to use by wireless service providers.

3.2. Network-Based Wireless Location

Network-based location technology depends on using the current cellular network to obtain wireless user location information. In these systems, the base stations (BSs) measure the signals transmitted from the MS and relay them to a central site for processing and calculating the MS location. The central processing site then relays the MS location information to the associated PSAP, as shown in Fig. 2. Such a technique is also known as *reverse-link wireless location*. Reverse-link wireless location has the main advantage of not requiring any modifications or specialized equipment in the MS handset, thus accommodating a large cluster of handsets already in use in existing cellular networks. The main disadvantage of network-based wireless location is its



Figure 2. Network-based wireless location.

relatively lower accuracy, when compared to GPS-based location methods [3].

Network-based wireless location techniques have the significant advantage that the MS is not involved in the location-finding process; thus these systems do not require any modifications to existing handsets. Moreover, they do not require the use of GPS components, thus avoiding any political issue that may arise from their use. However, unlike GPS location systems, many aspects of networkbased location are not fully studied yet. This is due to the relatively recent introduction of this technology. In most of the rest of this article, we will focus on networkbased wireless location. First, we will review the MS signal parameters that need to be estimated by the cellular base stations and how these signals are combined to obtain a MS location estimate, in data fusion, defined earlier. We will also discuss the sources of error that limit the accuracy of network-based location. Finally, we study different MS signal parameter estimation techniques along with some hardware implementation issues. Here, we may add that although many of the studied aspects apply to both GPSbased location and forward-link location, we will focus on reverse link network-based location. From this point on until the end of the article, we will refer to network-based wireless location simply as wireless location.

4. DATA FUSION METHODS

Data fusion for wireless location refers to combining signal parameter estimates obtained from different base stations to obtain an estimate of the MS location. We will study the conventional data fusion methods. The MS location coordinates in a Cartesian coordinate system are denoted by (x_0^o, y_0^o) , with the superscript 'o' used to denote quantities that are unknown and which we wish to estimate. These coordinates can be estimated from measured MS signal parameters, when measured at three or more base stations (BSs). The coordinates of the nearest three BSs to the MS, denoted by BS₁, BS₂, and BS₃, are $(x_1, y_1), (x_2, y_2)$, and (x_3, y_3) , respectively. Without loss of generality, the origin of the Cartesian coordinate system is set to those of BS₁:

$$(x_1, y_1) = (0, 0)$$

We will denote the time instant at which the MS starts transmission as time instant t_0^o . This MS signal reaches the three BSs involved in the MS location process at instants t_1^o , t_2^o , and t_3^o , respectively. The amplitudes of arrival of the MS signal at the main and adjacent sectors of BS_i are respectively denoted by A_{i1}^o and A_{i2}^o , for $i = 1, 2, 3.^3$ Data fusion methods obtain estimates for the MS coordinates, say, (x_0, y_0) , by combining the MS signals through

$$(x_0, y_0) = g(t_0, t_i, A_{i1}, A_{i2})$$
(1)

 3 In cellular systems, a sectored antenna structure is very common. Each BS usually contains three different antennas, with the main lobe of each antenna facing a different direction, and with an angle of 120° between each of the directions. The sector whose antenna faces a specific MS is termed the *main sector* serving this MS. The sector next to the main sector from the MS side is termed the *adjacent sector*.

where $\{t_0, t_i, A_{i1}, A_{i2}\}$ are estimates of $\{t_0^o, t_i^o, A_{i1}^o, A_{i2}^o\}$ and the function *g* depends on the data fusion method. The resulting location error from the data fusion operation is thus given by

$$e = \sqrt{(x_0 - x_0^o)^2 + (y_0 - y_0^o)^2}$$
(2)

One performance index, which is used to compare the accuracy of data fusion methods, is the location mean-square error (MSE), defined by

$$MSE = Ee^{2} = E[(x_{0} - x_{0}^{o})^{2} + (y_{0} - y_{0}^{o})^{2}]$$
(3)

Another performance index for data fusion methods is the value below which the error magnitude, |e|, lies for 67% of the time. In other words, it is the value of the error, $e_{67\%}$, at which the error cumulative density function (CDF) is equal to 0.67. The 67% error limit is the performance index that is used by the FCC to set the required location accuracy. Here we may add that for zero-mean Gaussian errors of variance σ_e^2 , we have

$$e_{67\%} = \sigma_e = \sqrt{\text{MSE}} \tag{4}$$

Several wireless location data fusion techniques have been introduced since the late 1990s, all of which are based on combining estimates of the time and/or amplitude of arrival of the MS signal when received at various BSs. These methods fall into the following categories:

- Time of arrival (ToA)
- Time difference of arrival (TDoA)
- Angle of arrival (AoA)
- Hybrid techniques

4.1. Time of Arrival (ToA)

The time of arrival (ToA) data fusion method is based on combining estimates of the time of arrival of the MS signal, when arriving at three different BSs. Since the wireless signal travels at the speed of light (C), thus the actual distance between the MS and BS_i , r_i , is given by

$$r_i^o = (t_i^o - t_0^o)C (5)$$

where t_0^o is the actual time instant at which the MS starts transmission and t_i^o is the actual time of arrival of the MS signal at BS_i. Each ToA estimate, t_i , serves to form an estimate of the distance between the MS and the corresponding BS as

$$r_i = (t_i - t_0)C \tag{6}$$

These estimated distances between the MS and each of the three BSs are then used to obtain (x_0, y_0) by solving the following set of equations:

$$r_1^2 = x_0^2 + y_0^2 \tag{7}$$

$$r_2^2 = (x_2 - x_0)^2 + (y_2 - y_0)^2 \tag{8}$$

$$r_3^2 = (x_3 - x_0)^2 + (y_3 - y_0)^2 \tag{9}$$



Figure 3. Some wireless location techniques: (a) TOA and (b) AOA. The MS is positioned at the intersection of the loci.

Without loss of generality, it can be assumed that $r_1 < r_2 < r_3$.

Now, a conventional way of solving this overdetermined nonlinear system of equations is as follows. First, equations (7) and (8) are solved for the two unknowns (x_0, y_0) to yield two solutions. As shown in Fig. 3a, each equation defines a locus on which the MS must lie. Second, the distance between each of the two solutions and the circle, whose equation is given by (9) is calculated. Finally, the solution that results in the shortest distance from the circle (9) is chosen to be an estimate of the MS location coordinates [4].

Although this method will help resolve the ambiguity between the two solutions resulting from solving Eqs. (7) and (8), it does not combine the third measurement r_3 in an optimal way. Furthermore, it is not possible to combine more ToA measurements from BSs more than three.

This can be solved by combining all the available set of measurements using a least-squares approach into a more accurate estimate. This approach can be summarized as follows. Subtracting (7) from (8), we obtain

$$r_2^2 - r_1^2 = x_2^2 - 2x_2x_0 + y_2^2 - 2y_2y_0$$

Similarly, subtracting (7) from (9), we obtain

$$r_{3}^{2} - r_{1}^{3} = x_{3}^{2} - 2x_{3}x_{0} + y_{3}^{2} - 2y_{3}y_{0}$$

Rearranging terms, the previous two equations can be written in matrix form as

$$\begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} K_2^2 - r_2^2 + r_1^2 \\ K_3^2 - r_3^2 + r_1^2 \end{bmatrix}$$
(10)

where

$$K_i^2 = x_i^2 + y_i^2 \tag{11}$$

Equation (10) can be rewritten as

r

$$\boldsymbol{H}\boldsymbol{x} = \boldsymbol{b} \tag{12}$$

where

$$\boldsymbol{H} = \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \boldsymbol{b} = \frac{1}{2} \begin{bmatrix} K_2^2 - r_2^2 + r_1^2 \\ K_3^2 - r_2^3 + r_1^2 \end{bmatrix}$$

The solution of (12) is given by

 $\boldsymbol{x} = \boldsymbol{H}^{-1}\boldsymbol{b}$

If more than three ToA measurements are available, it can be verified that (12) still holds, with

$$\boldsymbol{H} = \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \\ \vdots & \vdots \end{bmatrix}, \, \boldsymbol{b} = \frac{1}{2} \begin{bmatrix} K_2^2 - r_2^2 + r_1^2 \\ K_3^2 - r_3^2 + r_1^2 \\ K_4^2 - r_4^2 + r_1^2 \\ \vdots \end{bmatrix}$$

In this case, the least-squares solution of (12) is given by

$$\boldsymbol{x} = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{b}$$
(13)

The ToA method requires accurate synchronization between the BSs and MS clocks. Many of the current wireless system standards only mandate tight timing synchronization among BSs [e.g., 17]. However, the MS clock might have a drift that can reach a few microseconds. This drift directly reflects into an error in the location estimate of the ToA method.

4.2. Time Difference of Arrival (TDoA)

Another widely used technique that avoids the need for MS clock synchronization is based on time difference of arrival (TDoA) of the MS signal at two BSs. Each TDoA measurement forms a hyperbolic locus for the MS. Combining two or more TDoA measurements results in a MS location estimate that avoids MS clock synchronization errors [e.g., 18–21].

We now illustrate how a closed-form location solution can be obtained from TDoA measurements in the case of three BSs involved in the MS location. The TDoA measurement between BS_i and BS_1 is defined by

$$r_{i,1} \stackrel{=}{=} r_i - r_1$$

= $(t_i - t_0)C - (t_1 - t_0)C = (t_i - t_1)C$ (14)

Note that TDoA measurements are not affected by errors in the MS clock time (t_0) as it cancels out when subtracting two ToA measurements. Equation (8) can be rewritten, in terms of the TDoA measurement $r_{2,1}$, as

$$(r_{2,1}+r_1)^2 = K_2^2 - 2x_2x_0 - 2y_2y_0 + r_1^2$$

Expanding and rearranging terms, we get

$$-x_2x_0 - y_2y_0 = r_{2,1}r_1 + \frac{1}{2}(r_{2,1}^2 - K_2^2)$$

Similarly, we can write

Λ

$$-x_3x_0 - y_3y_0 = r_{3,1}r_1 + \frac{1}{2}(r_{3,1}^2 - K_3^2)$$

Rewriting these equations in matrix form we get

$$\boldsymbol{H}\boldsymbol{x} = \boldsymbol{c}\boldsymbol{r}_1 + \boldsymbol{d} \tag{15}$$

where

$$m{c} = egin{bmatrix} -r_{2,1} \ -r_{3,1} \end{bmatrix}, m{d} = rac{1}{2} egin{bmatrix} K_2^2 - r_{2,1}^2 \ K_3^2 - r_{3,1}^2 \end{bmatrix}$$

This equations can be used to solve for x, in terms of the unknown r_1 , to get

$$\boldsymbol{x} = \boldsymbol{H}^{-1}\boldsymbol{c}\boldsymbol{r}_1 + \boldsymbol{H}^{-1}\boldsymbol{d}$$

Substituting this intermediate result into (7), we obtain a quadratic equation in r_1 . Substituting the positive root back into the above equation yields the final solution for x.

If more than three BSs are involved in the MS location, Eq. (15) still holds with

$$\boldsymbol{H} = \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \\ \vdots & \vdots \end{bmatrix}, \boldsymbol{c} = \begin{bmatrix} -r_{2,1} \\ -r_{3,1} \\ -r_{4,1} \\ \vdots \end{bmatrix}, \boldsymbol{d} = \frac{1}{2} \begin{bmatrix} K_2^2 - r_{2,1}^2 \\ K_3^2 - r_{3,1}^2 \\ K_4^2 - r_{4,1}^2 \\ \vdots \end{bmatrix}$$

which yields the following least-squares intermediate solution

$$\boldsymbol{x} = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T (\boldsymbol{c} \boldsymbol{r}_1 + \boldsymbol{d})$$
(16)

Combining this intermediate result with (7), the final estimate for \boldsymbol{x} is obtained. A more accurate solution can be obtained as in Ref. 19 if the second-order statistics of the TDoA measurement errors are known.

4.3. Angle of Arrival (AoA)

In cellular systems, AoA estimates can be obtained by using antenna arrays. The direction of arrival of the MS signal can be calculated by measuring the phase difference between the antenna array elements or by measuring the power spectral density across the antenna array in what is known as beamforming (see, e.g., reference [22] and the works cited therein). Combining the AoA estimates of two BSs, an estimate of the MS position can be obtained (see Fig. 3b). Thus the number of BSs needed for the location process is less than that of ToA and TDoA methods by one. Another advantage of AoA location methods is that they do not need any BS clock synchronization. However, one disadvantage of using antenna-array-based location methods is that antenna array structures do not currently exist in second-generation (2G) cellular systems. Deploying antenna arrays in all existing BSs may lead to high cost burdens on wireless service providers. The use of antenna arrays is planned in some third-generation (3G) cellular systems, such as Universal Mobile Telecommunications System (UMTS) networks [e.g., 23,24], which will use antenna arrays to provide directional transmission in order to improve the network capacity.

AoA estimates can also be obtained using sectored multibeam antennas, which already exist in current cellular systems, using the technique described in reference [25]. In this technique, an estimate of the AoA $(\hat{\theta} - \text{see Fig. 4})$ is obtained based on the difference between the measured signal amplitude of arrival (AmpoA) at the main beam (beam 1) and the corresponding AmpoA

measured at the adjacent beam (beam 2).⁴ This difference is denoted by $A_1 - A_2$ in Fig. 5, where A_1 and A_2 are the measured amplitude levels in decibels. The measured AmpoA at the third beam may be used to resolve any ambiguity that might result from antenna sidelobes. One main challenge facing this technique is the relatively low signal-to-noise ratio (SNR) of the received MS signal at the adjacent beam, especially in cases where the AoA is close to a null in the adjacent beam field pattern (e.g., θ close to 0 degrees in Figs. 4 and 5). This significantly limits the AmpoA estimation accuracy at the adjacent beam.

4.4. Hybrid Techniques

In ToA, TDoA, and AoA methods, two or more BSs are involved in the MS location process. In situations where



Figure 4. Sectored-antenna field pattern.



Figure 5. Measured AmpOA level patterns (in dB) for a three-beam antenna versus the AOA $(\theta).$

 4 Here, *main beam* denotes the beam with the highest received signal level and *adjacent beam* refers to the beam that receives the second highest signal level.

the MS is much closer to one BS (serving site) than the other BSs, the accuracy of these methods is significantly degraded because of the relatively low SNR of the received MS signal at one or more BSs. Such accuracy is further reduced due to the use of power control, which requires the MS to reduce its transmitted power when it approaches a BS, causing what is known as the *hearability* problem [26]. Such problems will be discussed in the next section. In these cases, an alternate location procedure is to obtain an angle of arrival estimate (AoA) from the serving site and combine it with a ToA estimate of the serving site [27]. Combining ToA and AoA estimates from one BS leads to one well-defined MS position estimate, which corresponds to the intersection of a circle and a straight line that starts at the center of the circle. The precision of this hybrid technique is limited by the accuracy of the ToA measurement, which is dictated by the accuracy of the MS clock. Many other hybrid location data fusion techniques can be used, such as combining TDoA and AoA measurements [28].

5. SIGNAL PARAMETER ESTIMATION

From the previous discussion, we can see that the wireless location methods depend on combining estimates of the ToA and/or AoA of the received signal at/from different BSs. Although estimating the time and amplitude of arrival of wireless signals has been studied in many works since 1990 as it is needed in many cellular systems for online signal decoding purposes [29], parameter estimation for wireless location is actually a different estimation problem in many respects. This makes the success of using conventional estimation algorithms very limited in wireless location problems. In this section, we will illustrate the differences between signal parameter estimation for conventional signal decoding and wireless location. We will then discuss some particular system issues that makes signal estimation for wireless location different from one cellular system to the other (e.g., GSM, 2G and 3G CDMA systems).

Signal parameter estimation for wireless location purposes is different than that for online signal decoding in the following aspects:

1. Lower SNRs. Cellular systems usually suffer from high multiple-access interference levels that degrade the SNR of the received signal, thus degrading the signal parameter estimation accuracy in general. Moreover, for network-based wireless locations, the ability to detect the MS signal at multiple base stations is limited by the use of power control algorithms, which require the MS to decrease its transmitted power when it approaches the serving BS. This significantly decreases the received MS signal power level, when received at other BSs involved in the location process. This scenario is shown in Fig. 6, where the received SNRs at BS_1 and BS_2 are significantly reduced as the target MS approaches BS3. In a typical CDMA IS95 cellular environment, the received SNR of the serving BS is in the order of -15 dB. Conventional signal estimation algorithms are usually designed to work at this SNR level. However, the received SNR at BSs other than



Figure 6. Multiple-access interference among adjacent cells in cellular systems. The letters BS indicate the base stations and the letters MS denote the target mobile station.

the main serving BS can be as low as -40 dB, which poses a challenge for wireless location in such environments.

2. Almost Perfect Knowledge of Transmitted Signals. In conventional signal parameter estimation for online signal decoding, the transmitted MS bits are unknown. This forces signal estimation algorithms to perform a squaring operation to remove any bit ambiguity. The squaring operation limits the period over which coherent signal integration (averaging) is possible to the bit period. Further signal integration is only possible in a noncoherent manner, that is averaging after squaring. In wireless location applications, signal estimation algorithms can have almost perfect knowledge of the MS signal in many cases. For example, at the serving site, the MS signal is decoded with reasonably high accuracy (within a 1% frame error rate). The decoded bits become ready for use after a delay that is equal to the decoded frame period used in the cellular system (20 ms for IS95 systems). Because of the nature of wireless location applications, such a delay is not critical. Thus, the received MS signal can be buffered or delayed until the decoded bits become available through the conventional decoding process. Moreover, in many cellular systems, a cyclic redundancy check (CRC) feature is used. This enables the decoder to point out the erroneous frames after the decoding process. These erroneous frames can be ignored in the signal estimation process. The decoded bit information, obtained from the main sector of the serving site, can also be used by other adjacent sectors of the same site. Furthermore, this bit information can be transmitted through the network infrastructure to other BSs involved in locating the MS. This is known as tape recording of the MS signal. Another technique that avoids the tape recording process is known as the powerup function (PuF), which requires the MS in emergency situations to override the power control commands and raise its transmitted power level above the conventional level. Moreover, the MS transmits known probing bit sequences instead of its regular unknown bit sequence for a part or all of the transmission period. Although this solution overcomes many of the difficulties

encountered at far BSs, it requires modifying the existing handsets or at least the used power control algorithms. Furthermore, it can cause a decrease in the overall network capacity [26].

3. Channel Fading. Channel fading is considered constant during the relatively short estimation period of conventional signal parameter algorithms for online signal decoding, and is thus ignored in the design of such algorithms. This assumption cannot be made for wireless location applications where the estimation period could be considerably longer (might reach a few seconds). Furthermore, coherent integration periods are no longer limited by the bit duration, much longer coherent averaging periods could be achieved in wireless location applications [30,31]. In this case, the coherent integration period is limited by the received signal phase rotation. Thus, unlike the case of online channel estimators, channel fading plays an important role in any successful design of signal parameter estimators for wireless location. In many cases, the system parameters have to be adapted to the available knowledge of the channel fading characteristics.

4. Need to Resolve Overlapping Multipath. Multipath propagation is often encountered in wireless channels (see, e.g., the paper [32] and the references cited therein). In wireless location systems, the accurate estimation of the time and amplitude of arrival of the first arriving ray of the multipath channel is vital. In general, the first arriving (prompt) ray is assumed to correspond to the most direct path between the MS and the BS. However, in many wireless propagation scenarios, the prompt ray is succeeded by a multipath component that arrives at the receiver within a short delay from the prompt ray. If this delay is smaller than the duration of the pulse-shape used in the wireless system, these two rays overlap causing significant errors in the prompt ray time and amplitude of arrival estimation. Resolving these overlapping multipath components becomes rather difficult in low SNR and rapid channel fading situations. On the other hand, resolving these overlapping components is not vital for signal decoding applications as it does not significantly affect the performance of the signal decoding operation, for which coarse estimates for the channel time delays and amplitudes are sufficient.

Figure 7 shows an example for the combined impulse response of a two-ray channel and a conventional pulseshape for a conventional CDMA IS95 system in two cases (a,b). In case (a), the delay between the two channel rays is equal to twice the chip duration $(2T_c)$. It is clear that the peaks of both rays are resolvable, by a simple peak picking procedure, thus allowing for relatively accurate estimation of the prompt ray time and the amplitude of arrival. However, in case (b), both multipath components overlap and are *nonresolvable* via peak picking. This can lead to significant errors in the prompt ray time and amplitude of arrival estimation. These errors cannot be tolerated for wireless location applications, especially in the case of a relatively wide pulseshaping waveform.

5.1. Parameter Estimation Schemes

We now elaborate on some schemes that are used to estimate the wireless signal time and amplitude of arrival.



Figure 7. Overlapping rays: (a) delay = $2T_c$; (b) delay = $T_c/2$.

The aim of such schemes is to estimate an unknown constant discrete-time delay, τ^{o} , of a known real-valued sequence $\{s(n)\}$. The signal is transmitted over a single-path time-varying channel, and the designer has access to a measured sequence $\{r(n)\}_{n=1}^{K}$ that relates to $\{s(n)\}$ via

$$r(n) = Ax^{o}(n)s(n - \tau^{o}) + v(n)$$
(17)

where v(n) is additive white Gaussian noise, and $\{x^o(n)\}\$ accounts for the time-varying nature of the *fading* channel gain over which the sequence $\{s(n)\}\$ is transmitted, while A is a constant *unknown* received signal amplitude that accounts for both the gain of the *static* channel if fading were not present and the antenna beam gain. Multipath issues are considered later in this section.

A conventional estimation scheme for τ^o for online bit decoding purposes is shown in Fig. 8. In this scheme, the received sequence, r(n), is correlated with replicas of $\{s(n - \tau_i)\}$ over a grid of τ values, say, $\{\tau_1, \tau_2, \ldots, \tau_F\}$. The coherent averaging period, N, is set to the bit interval. The outputs of the correlation process are squared to remove any bit ambiguity and then noncoherently averaged over the rest of the available estimation period.



Figure 8. Conventional time-delay estimation for single-path channels.

Figure 9 shows a block diagram of a wireless location ToA/AoA estimation scheme [30]. In this scheme, the received sequence $\{r(n)\}$ is also multiplied by a replica of the transmitted sequence $\{s(n - \tau)\}$ for different values of τ . The resulting sequence is then averaged coherently over an interval of N samples, and further averaged noncoherently for M samples to build a power delay profile, $J(\tau)$. The averaging intervals N and M are positive integers that satisfy K = NM, and the value of N is picked adaptively in an optimal manner by using an estimate of the maximum Doppler frequency of the fading channel (\hat{f}_D) , which can be estimated using some suggested techniques [e.g., 33].

The searcher picks the maximum of $J(\tau)$, which is given by

$$J(\tau) = \frac{1}{M} \sum_{m=1}^{M} \left| \frac{1}{N} \sum_{n=(m-1)N+1}^{mN} r(n) s(n-\tau) \right|^2$$
(18)

and assigns its index to the ToA estimate, according to

$$\widehat{\tau^o} = \arg\max J(\tau) \tag{19}$$

The optimal value of the coherent averaging period (N_{opt}) is obtained by maximizing the SNR gain at the output of the estimation scheme with respect to N which leads to [30]

$$\sum_{i=1}^{N_{opt}-1} iR_x(i) = 0$$
 (20)

where $R_x(i)$ is the autocorrelation function of the sequence $\{x(n)\}$. For a Rayleigh fading channel, $R_x(i)$ is given by

$$R_x(|i|) = J_0(2\pi f_D T_s i)$$

where $J_0(\cdot)$ is the first-order Bessel function, T_s is the sampling period of the received sequence $\{r(n)\}$, and f_D is



Figure 9. A time-delay estimation scheme for single-path fading channels.

the maximum Doppler frequency of the Rayleigh fading channel. Equation (20) shows that the coherent averaging interval N should be adapted according to the channel autocorrelation function.

It has been shown [30,31] that when coherent/noncoherent averaging estimation schemes are used for wireless location applications, where an extended coherent averaging interval is used, two biases arise at the output of the estimation scheme. Both biases affect the accuracy of the amplitude estimate significantly. The first bias is an additive noise bias that increases with the noise variance and is given by

$$B_n = \frac{\sigma_v^2}{N} \tag{21}$$

The second bias is a multiplicative fading bias that depends on the autocorrelation function and is given by

$$B_f = \frac{R_x(0)}{N} + \sum_{i=1}^{N-1} \frac{2(N-i)R_x(i)}{N^2}$$
(22)

It is clear that B_f is less than or equal to unity (it is unity for static channels, which explains why previous conventional designs ignored this bias as fading was not considered in these designs [29]; the value of B_f is also unity for N = 1).

To correct for these biases, the searcher equalizes the peak value of $J(\tau)$ by subtracting two fading and noise biases, which are estimated by means of the upper and lower branches of the scheme of Fig. 9. The output of this correction procedure is taken as an estimate for the amplitude of arrival, which is given by

$$A = \sqrt{C_f[J(\tau^o) - B_n]}$$

The value of C_f (the fading correction factor) is $C_f = 1/B_f$:

$$C_f = \left[\frac{R_x(0)}{N} + \sum_{i=1}^{N-1} \frac{2(N-i)R_x(i)}{N^2}\right]^{-1}$$
(23)

For a Rayleigh fading channel, this correction factor increases with the maximum Doppler frequency of the fading channel. When f_D is estimated, we actually end up

with an estimate for C_f . For the case of CDMA systems, the quantity B_n can be estimated as follows. Note first that the noise variance σ_v^2 can be estimated *directly* from the received sequence $\{r(n)\}$ since, for CDMA signals, the SNR is typically very low. In other words, we can get an estimate for σ_v^2 as follows:

$$\widehat{\sigma_v^2} = \frac{1}{K} \sum_{i=1}^K |r(i)|^2$$

Then, an estimate for B_n is given, from (21), by

$$\widehat{B}_{n} = \frac{\widehat{\sigma_{v}^{2}}}{N} = \frac{1}{NK} \sum_{i=1}^{K} |r(i)|^{2}$$
(24)

With $\{B_n, C_f\}$ so computed, we obtain an estimate for A via the expression

$$\widehat{A} = \sqrt{C_f[J(\widehat{\tau^o}) - \widehat{B}_n]}$$
(25)

More details on this scheme and simulation results can be found in the literature [30,31,34].

5.2. Overlapping Multipath Resolving

As mentioned before, wireless propagation usually suffers from severe multipath conditions. In situations where the prompt ray overlaps with a successive ray, a significant error in both the time and amplitude of arrival estimation is encountered.

Overlapping multipath components can be modeled by considering the relation

$$r(n) = c(n) * p(n) * h(n) + v(n)$$
(26)

where $\{r(n)\}$ continues to denote the received sequence, $\{c(n)\}$ is a known binary sequence, $\{p(n)\}$ is a known pulseshape impulse response sequence, v(n) is additive white Gaussian noise of variance σ_v^2 , and h(n) now refers to a multipath channel that is described by

$$h(n) = \sum_{l=1}^{L} \alpha_l \, x_l(n) \delta(n - \tau_l^o) \tag{27}$$

Here α_l , $\{x_l(n)\}$, and τ_l^o are respectively the unknown gain, the normalized amplitude sequence, and the time of arrival of the *l*th multipath component (ray). The above model assumes that there is a multipath component at each delay with corresponding amplitude α_l . In practice, most of these amplitudes will be zero or insignificant. For this reason, a common procedure is to estimate the amplitudes at all delays and to compare them to a threshold value that is proportional to the noise variance. If the amplitude α_l , at a specific delay τ_l^o , is larger than the threshold, then it is declared to correspond to a multipath component. The time and amplitude of arrival are then taken as the time and amplitude of the earliest ray higher than this threshold.

In this regard, the required estimation problem is one of estimating the vector of amplitudes at all possible delays, which is given by

$$\boldsymbol{h} \stackrel{\Delta}{=} \operatorname{col}[\alpha_1, \alpha_2, \ldots, \alpha_L]$$

Several least-squares-type methods have been suggested for this purpose [35–37]. These methods exploit the known transmitted pulse-shape to resolve overlapping rays. For example, it has been shown [37] that, under some reasonable assumptions, the vector \boldsymbol{h} can be estimated by means of the following procedure. The received sequence is multiplied by delayed replica of the known transmitted sequence, $\{s(n - \tau)\}$. Each N sample of the resulting sequence is coherently averaged and the resulting averages at all delays are collected into a vector, say, \boldsymbol{r} . An estimate of \boldsymbol{h} is then obtained from \boldsymbol{r} by solving a least-squares problem, which leads to

$$\hat{\boldsymbol{h}} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{r}$$
(28)

where A denotes a convolution matrix that is constructed from the pulse-shaping waveform. A general block diagram for such least-squares based techniques is shown in Fig.10. Alternative so-called *superresolution* techniques are also available that are based on methods known as ESPRIT and MUSIC (see the paper [22] and references cited therein).

Least-squares multipath resolving techniques, however, suffer from noise boosting, which is usually caused by the ill conditioning of the matrices involved in the LS operation. This ill-conditioning magnifies the noise at the output of the LS stage. For wireless location finding applications, where the received signal-to-noise ratio (SNR) is relatively low, noise magnification leads to significant errors in the time and amplitude of arrival estimates, which in turn result in low location precision. Other modified LS techniques that attempt to avoid matrix ill conditioning—such as regularized least-squares, total least-squares, and singular value decomposition methods—lack the required fidelity to resolve overlapping multipath components. Furthermore, applying least-squares methods may produce unnecessary errors in the case of single-path propagation.

An adaptive filtering technique for multipath resolving that avoids the aforementioned difficulties has been discussed [38]. Although adaptive filters do not suffer from noise amplification, they can still suffer from slow convergence and also divergence in some cases. These problems can be addressed by using knowledge about the channel autocorrelation and the fact that each channel ray fades at a different Doppler frequency.

6. HARDWARE IMPLEMENTATION ISSUES

It is clear from the previous considerations that signal parameter estimation for wireless location purposes often requires performing an extensive search over a dense grid of the estimated parameter (e.g., ToA estimation). The hardware implementation of these search schemes requires special attention as they might introduce a dramatic increase in the overall system hardware complexity and power consumption. In this section we review two hardware architectures for implementing ToA estimation schemes. The first scheme depends on combing the hardware of both channel and location searchers, while the second involves a Fast Fourier Transform (FFT)-based estimation scheme. Both architectures aim at reducing the overall hardware complexity.

6.1. Combined Channel/Location Searchers

A main hardware block in CDMA receivers is the conventional RAKE receiver, which consists of a dedicated channel searcher and a minimum of three RAKE fingers. Channel searchers obtain coarse estimates of the time and amplitude of arrival of the strongest multipath components of the MS signal. This information is then used by the receiver RAKE fingers and delay-locked loops (DLLs) to lock onto the strongest channel multipath components, which are combined and used in bit decoding. Estimates of the time and amplitude of arrival of the strongest rays are continuously fed from the channel searcher to the RAKE fingers.

Although the location searcher and RAKE receiver differ in purpose, structure, and estimation period, several basic building hardware blocks used in each of them are common. This fact can be exploited to combine both searchers into a single architecture that serves to save hardware blocks with added design flexibility.

Figure 11 shows the scheme, proposed in another paper [39], for the combined searcher architecture. The



Figure 10. Multipath least-squares searcher.



Figure 11. Combined architecture for location searcher and RAKE receiver.

scheme is formed from L_l data branches. Each data branch starts with a correlator over N_1 samples (despreader), where N_1 is the number of chips per symbol multiplied by the number of data samples per chip (4, 8, or 16). The output of the correlator is then multiplexed between two paths, marked "1" and "2" in Fig. 11. In path 1, which corresponds to data path of the channel searcher or RAKE finger branches, the despread signal is squared and noncoherently averaged over M_1 samples, where M_1 is optionally adapted to an estimate of the maximum Doppler frequency of the fading channel. For path 2, which is needed for the location parameter estimation, the despread sequence is delayed for a frame period, multiplied by an estimate of the transmitted bit sequence, coherently averaged over N_2 samples, squared, and noncoherently averaged over M_2 samples. Both N_2 and M_2 are adapted according to an estimate of the maximum Doppler frequency. The output of either paths is used to extract the channel multipath parameters.

The dynamic operation of the scheme is as follows. The received sequence is despread by multiplying by delayed code replica $c(n - \tau)$ and averaged over N_1 samples after which the N_1 register is reset. For online bit decoding, samples of the despread sequence are squared and noncoherently averaged. The average of every M_1 samples is passed to the multipath parameter extraction block and the M_1 register is then reset. Coarse multipath rays information are continuously fed to the L_f RAKE fingers, which use $3L_f$ branches of the scheme to obtain early, on-time, and late correlations over M_1 symbols. Such correlations are needed to advance or delay the sampling timing to lock onto the correct sampling point. This is done according to the difference between the early and late correlations [40]. The outputs of these fingers are combined, and used in bit decoding. Optional DLLs can be used to further enhance the tracking performance of the RAKE fingers. For location parameter estimation, the despread sequence is delayed, multiplied by an estimate of the transmitted bit sequence $\hat{b}(n)$, and continuously averaged over N_2 symbols. Every N_2 symbols, the N_2 register is reset and its output is squared using the shared squaring circuits and averaged over M_2 samples. After the total location estimation period ($N_2 \times M_2$ symbols), the time and amplitude of arrival of the prompt ray are equalized for fading and noise biases and used to extract needed location parameters.

This architecture has the following advantages:

- 1. Saving a large number of hardware building blocks via multiplexing basic hardware blocks between $3L_f + L_c$ location searcher and RAKE receiver branches.
- 2. Improving the performance of the RAKE receiver by continuously adapting the estimation period M_1 to an estimate of the maximum Doppler frequency. This period is conventionally adjusted to track a fading channel in the worst (fastest) case, which restricts this period to a small value (around 6 symbols for IS95 systems). Adapting the estimation period of the channel searcher has two advantages: (a) it will increase the accuracy of the delay and amplitude estimates, and (b) it will help save power as it will reduce the number of times the RAKE fingers need to change their lock point, especially for low maximum Doppler frequency cases.
- 3. Reducing the hardware complexity significantly by eliminating the need to use DLLs for fine tracking in the cases where the accuracy of the used combined architecture is $T_c/8$ or higher (which is typical for location applications). In such cases the accuracy of the RAKE receiver will be adequate for online bit decoding without the use of DLLs. Hardware implementation of DLLs is extremely complex, especially with regard to the analog front end [40].

Further details of the operation of this architecture is given in [39], along with performance simulation results.

6.2. FFT-Based Searchers

In mobile-based wireless location systems, a maximumlikelihood searcher is embedded in the MS handset. It is very common for such searchers to involve multiple correlation operations of the received signals, from cellular BSs or GPS satellites, with local delayed replica of the transmitted signals. Performing these extensive correlations in the time domain may be a burden for the MS hardware. Often these correlations are performed using a general-purpose DSP processor, which is embedded in the MS to perform many other tasks, including the correlation process. DSP processors have many advantages, such as low cost, versatility, and design flexibility. An efficient way of implementing correlation operations in this case is through the use of fast Fourier transform (FFT). FFTbased location searchers have shown significant efficiency when implemented on DSP processors [e.g., 15,41].

We now review the basic principles of operation of FFTbased location searchers. The output of the correlation operation, say, $y(\tau)$, between two sequences, $\{r(n)\}$ and $\{s(n)\}$, can be viewed as a sum of the form

$$y(\tau) = \sum_{n=1}^{N} r(n)s(n-\tau)$$

Evaluating this sum requires N multiplication operations for every value of the delay τ . Thus, computing $y(\tau)$ in the time domain needs N^2 multiplications. On the other hand, the correlation operation can be viewed as a multiplication of two sequences in the frequency domain, which has the form

$$Y(\omega) = R(\omega) \cdot S(\omega)$$

where $Y(\omega)$, $R(\omega)$, and $S(\omega)$ are the Fourier transforms of $y(\tau)$, r(n), and s(n), respectively. Thus, an alternate way of computing $y(\tau)$, is via

$$y(\tau) = \mathcal{F}^{-1}[R(\omega) \cdot S(\omega)].$$

Thus, the total number of multiplications needed to perform this operation is the number of multiplications needed to obtain the FFT of the sequences r(n) and s(n), the product of $R(\omega)$ and $S(\omega)$, and finally the inverse FFT of this product. Notice that if the sequence s(n) is perfectly known, its FFT can also be known and stored instead of storing the signal s(n) itself. Thus the total number of multiplications needed is given by $(N + N \log_2 N)$. When N is relatively large, this number of multiplications can be significantly less than N^2 . For example, correlating GPS signals of length 1024 using the FFT approach is faster than performing the correlation in the time domain by a factor of 64 [15,41]. This directly reflects to a huge saving in the complexity and power consumption of the MS handset. We may also note that this procedure works only for sequences whose length is a power of 2. The general case can still be efficiently treated using the chirpz transform (CZT), which can handle sequences whose length is not a power of two (See Refs. 15 and 41 for more details).

7. CONCLUDING REMARKS

As can be inferred from the discussions in the body of the article, and from the extended list of references below, wireless location is an active field of investigation with many open issues and with a variety of possible approaches and techniques. The final word is yet to come, which opens the road to much further work and, ultimately, to tremendous benefits.

Acknowledgments

This work was partially supported by the National Science Foundation under Award numbers CCR-9732376 and ECS-9820765. The authors would also like to thank Dr. L. M. A. Jalloul for his input, insights and collaboration in this area of research.

BIOGRAPHIES

Ali H. Sayed received his B.S. and M.S. degrees in electrical engineering from the University of Sao Paulo, Brazil, and his Ph.D. degree in electrical engineering in 1992 from Stanford University. He is professor of electrical engineering at the University of California, Los Angeles. He has over 170 publications, is the coauthor of two published books and the coeditor of a third book. He sits on the editorial boards of several journals including the IEEE Transactions on Signal Processing, the SIAM Journal on Matrix Analysis and Its Applications, and the International Journal of Adaptive Control and Signal Processing. He is also a member of the technical committees on signal processing theory and methods (SPTM) and on signal processing for communications (SPCOM), both of the IEEE Signal Processing Society. He has contributed several articles to engineering and mathematical encyclopedias and handbooks, and has served on the program committees of several international meetings. He has also consulted with industry in the areas of adaptive filtering, adaptive equalization, and echo cancellation. His research interests span several areas including adaptive and statistical signal processing, filtering and estimation theories, equalization techniques for communications, interplays between signal processing and control methodologies, and fast algorithms for largescale problems. Dr. Sayed is a recipient of the 1996 IEEE Donald G. Fink Award and is a fellow of IEEE.

Nabil R. Yousef received his B.S. and M.S. degrees in electrical engineering from Ain Shams University, Cairo, Egypt, in 1994 and 1997, respectively, and his Ph.D. in electrical engineering from the University of California, Los Angeles, in 2001. He is currently a senior research staff member at Broadcom Corp., Irvine, California. His research interests include adaptive filtering, equalization, CDMA systems, and wireless location. He is a recipient of a 1999 Best Student Paper Award at an international meeting for work on adaptive filtering, and of a 1999 NOKIA Fellowship Award.

BIBLIOGRAPHY

 FCC Docket 94-102, Revision of the Commissions Rules to Insure Compatibility with Enhanced 911 Emergency Calling Systems, Technical Report RM-8143, July 1996.

- State of New Jersey, Report on the New Jersey Wireless Enhanced 911 Terms: The First 100 Days, Technical Report, June 1997.
- J. H. Reed, K. J. Krizman, B. D. Woerner, and T. S. Rappaport, An overview of the challenges and progress in meeting the E-911 requirement for location service, *IEEE Commun. Mag.* 36(4): 30-37 (April 1998).
- J. J. Caffery and G. L. Stuber, Overview of radiolocation in CDMA cellular systems, *IEEE Commun. Mag.* 36(4): 38-45 (April 1998).
- J. J. Caffery and G. L. Stuber, Radio location in urban CDMA microcells, *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Toronto, Canada, Sept. 1995, Vol. 2, pp. 858–862.
- J. M. Zagami, S. A. Parl, J. J. Bussgang, and K. D. Melillo, Providing universal location services using a wireless E911 location network, *IEEE Commun. Mag.* 36(4): 66-71 (April 1998).
- T. S. Rappaport, J. H. Reed, and B. D. Woerner, Position location using wireless communications on highways of the future, *IEEE Commun. Mag.* 34(10): 33-41 (Oct. 1996).
- L. A. Stilp, Carrier and end-user application for wireless location systems, *Proc. SPIE*, Philadelphia, Oct. 1996, Vol. 2602, pp. 119–126.
- I. Paton et al., Terminal self-location in mobile radio systems, Proc. 6th Int. Conf. Mobile Radio and Personal Communications, Coventry, UK, Dec. 1991, Vol. 1, pp. 203-207.
- A. Giordano, M. Chan, and H. Habal, A novel locationbased service and architecture, *Proc. IEEE PIMRC*, Toronto, Canada, Sept. 1995, Vol. 2, pp. 853–857.
- J. J. Caffery and G. L. Stuber, Vehicle location and tracking for IVHS in CDMA microcells, *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Amsterdam, Netherlands, Sept. 1994, Vol. 4, pp. 1227–1231.
- B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins, Global Positioning System: Theory and Practice, 2nd ed., Springer-Verlag, New York, 1993.
- 13. Special issue on GPS, Proc. IEEE 87: 3-172 (Jan. 1999).
- S. Chakrabarti and S. Mishra, A network architecture for global wireless position location services, *Proc. IEEE Int. Conf. Communications*, Vancouver, BC, Canada, June 1999, Vol. 3, pp. 1779–1783.
- 15. U.S. Patent 5,663,734 (Sept., 1997), N. Krasner, GPS receiver and method for processing GPS signals.
- P. K. Enge, R. M. Kalafus, and M. F. Ruane, Differential operation of the Global Positioning System, *IEEE Commun. Mag.* 26(7): 48-60 (July 1988).
- 17. Telecommunications Industry Association, The CDMA2000 ITU-R RTT Candidate Submission VO. 18, July 1998.
- K. C. Ho and Y. T. Chan, Solution and performance analysis of geolocation by TDOA, *IEEE Trans. Aerospace Electron.* Syst. 29(4): 1311-1322 (Oct. 1993).
- Y. T. Chan and K. C. Ho, A simple and efficient estimator for hyperbolic location, *IEEE Trans. Signal Process.* 42(8): 1905-1915 (Aug. 1994).
- R. Schmidt, Least squares range difference location, IEEE Trans. Aerospace Electron. Syst. 32(1): 234-242 (Jan. 1996).
- B. T. Fang, Simple solutions for hyperbolic and related position fixes, *IEEE Trans. Aerospace Electron. Syst.* 26(5): 748-753 (Sept. 1990).

- H. Krim and M. Viberg, Two decades of array signal processing research: The parametric approach, *IEEE Signal Process. Mag.* 13(4): 67-94 (July 1996).
- T. Ojanpera and R. Prasad, Wideband CDMA for Third Generation Mobile Communications, Artech House, Boston, 1998.
- 24. R. Prasad, W. Mohr, and W. Konhauser, *Third Generation Mobile Communication Systems*, Artech House, Boston, 2000.
- K. Kuboi et al., Vehicle position estimates by multibeam antennas in multipath environments, *IEEE Trans. Vehic. Technol.* 41(1): 63-68 (Feb. 1992).
- A. Ghosh and R. Love, Mobile station location in a DS-CDMA system, *Proc. IEEE Vehicular Technology Conf.*, May 1998, Vol. 1, pp. 254–258.
- 27. K. J. Krizman, T. E. Biedka, and T. S. Rappaport, Wireless position location: fundamentals, implementation strategies, and sources of error, *Proc. IEEE Vehicular Technology Conf.*, New York, May 1997, Vol. 2, pp. 919–923.
- G. P. Yost and S. Panchapakesan, Automatic location identification using a hybrid technique, *Proc. IEEE Vehicular Technology Conf.*, May 1998, Vol. 1, pp. 264–267.
- S. Glisic and B. Vucetic, Spread Spectrum CDMA Systems for Wireless Communications, Artech House, Boston, 1997.
- N. R. Yousef and A. H. Sayed, A new adaptive estimation algorithm for wireless location finding systems, *Proc. Asilomar Conf.*, Pacific Grove, CA, Oct. 1999, Vol. 1, pp.491-495.
- N. R. Yousef, L. M. A. Jalloul, and A. H. Sayed, Robust timedelay and amplitude estimation for CDMA location finding, *Proc. IEEE Vehicular Technology Conf.*, Amsterdam, Netherlands, Sept. 1999, Vol. 4, pp. 2163–2167.
- R. Ertel et al., Overview of spatial channel models for antenna array communication systems, *IEEE Pers. Commun. Mag.* 5(1): 10-22 (Feb. 1998).
- A. Swindlehurst, A. Jakobsson, and P. Stoica, Subspacebased estimation of time delays and Doppler shifts, *IEEE Trans. Signal Process.* 46(9): 2472-2483 (Sept. 1998).
- G. Gutowski et al., Simulation results of CDMA location finding systems, *Proc. IEEE Vehicular Technology Conf.* Houston, TX, May 1999, Vol. 3, pp. 2124-2128.
- 35. Z. Kostic, M. I. Sezan, and E. L. Titlebaum, Estimation of the parameters of a multipath channel using set-theoretic deconvolution, *IEEE Trans. Commun.* 40(6): 1006-1011 (June 1992).
- 36. T. G. Manickam and R. J. Vaccaro, A non-iterative deconvolution method for estimating multipath channel responses, Proc. Int. Conf. Acoustics, Speech, and Signal Processing, April 1993, Vol. 1, pp. 333–336.
- N. R. Yousef and A. H. Sayed, Overlapping multipath resolving in fading conditions for mobile positioning systems, *Proc. IEEE 17th Radio Science Conf.*, Minuf, Egypt, Feb. 2000, Vol. 1, No. C-19, pp. 1–8.
- N. R. Yousef and A. H. Sayed, Adaptive multipath resolving for wireless location systems, *Proc. Asilomar Conference on Signals, Systems, and computers*, Pacific Grove, CA, Nov. 2001.
- 39. N. R. Yousef and A. H. Sayed, A new combined architecture for CDMA location searchers and RAKE receivers, *Proc. Int. Symp. Circuits and Systems*, Geneva, Switzerland, May 2000, Vol. 3, pp. 101–104.

2972 WIRELESS MPEG-4 VIDEOCOMMUNICATIONS

- 40. S. Sheng and R. Brodersen, *Low-Power CMOS Wireless Communications. A Wideband CDMA System Design*, Kluwer, Boston, 1998.
- R. G. Davenport, FFT processing of direct sequence spreading codes using modern DSP microprocessors, *Proc. IEEE Natl. Aerospace and Electronics Conf.*, Dayton, OH, May 1991, Vol. 1, No. 3, pp. 98–105.
- 42. T. Rappaport, Wireless Communications; Principles and Practice, Prentice-Hall, Englewood Cliffs, NJ, 1996.

WIRELESS MPEG-4 VIDEOCOMMUNICATIONS*

MADHUKAR BUDAGAVI Texas Instruments, Incorporated Dallas, Texas

1. INTRODUCTION

With the success of personal mobile wireless phones for voice communications, there is now wide commercial interest and activity in extending the capabilities of the mobile phone to support videocommunications. Addition of video functionality to mobile phones leads to several new applications of the mobile phone-these include videotelephony, streaming video, video e-postcards and messaging, surveillance, and distance learning and collaboration. Mobile videotelephony enables users to not only talk to each other anywhere and at any time they want to, but it also allows them to see each other at the same time. Streaming video turns the mobile phone into a mobile entertainment device — it enables users to watch news and sports clips, music videos, and movie clips at any place and at any time they want to. It also allows mobile phone users to watch the video streaming from their home camera for surveillance purposes. Support for sending video e-postcards and messages in mobile phones enables users, for example, to send their vacation videos and photos directly from their vacation spots itself. Support for receiving instant video messages enables users to be immediately notified of any security event detected on their home surveillance camera. Users can take a look at the video of the security event enclosed with the instant video message and decide what action to take. Users can also use their video-enabled mobile phones to look at real-time educational lectures and videos and get trained on their long commute to work on trains. They can also use the video-enabled mobile phone to remotely collaborate from a worksite; for instance, they can use their mobile videophone to send real-time images of ongoing construction to their colleagues in their office and collaborate. One can similarly think of many other applications of a video-enabled mobile phone or device.

Wireless videocommunications is a multifaceted problem covering the fields of signal processing, wireless communications, data compression, transport protocols, and microelectronics. Supporting video transmission on wireless channels involves many technical challenges. Raw digital video data require a large amount of bandwidth; for instance, even a low-resolution $(176 \times 144$ -pixel) color video sequence at 15 frames per second (fps) requires 4.5 megabits per second (Mbps). The bandwidth available on current wireless channels is limited and also expensive. Hence it becomes important that the video data be compressed prior to transmission over wireless channels. Video sequences have redundancies in both the temporal (i.e., between adjacent video frames) and the spatial (i.e., within a video frame) domains. Video compression is achieved by removing these redundancies. Standard video compression algorithms usually make use of the following three steps to achieve efficient compression:

- 1. Predict the current video frame from the previous video frame (by using motion vectors) to remove temporal redundancy.
- 2. Then use the energy-compacting discrete-cosine transform (DCT) to encode spatial redundancy.
- 3. Finally, use entropy coding (variable-length coding) to encode the various parameters resulting from steps 1 and 2.

Advances in low-bit-rate video coding now enable a 176×144 -pixel resolution color video sequence at 15 fps (which requires about 4.5 Mbps to be transmitted in the uncompressed form) to be compressed to about 32-64 kbps while still maintaining adequate viewing quality. The amount of compression that can be achieved is strongly dependent on the content in the video sequence. Video sequences with low motion can be usually compressed more efficiently than video sequences with high motion.

Current second-generation wireless systems provide data rates of only about 9.6-13 kbps. This amount of bandwidth is not enough for acceptable quality videocommunications. Advances in wireless technology and increased spectrum availability have lead to the development of third-generation (3G) wireless systems, which provide bandwidths of \leq 384 kbps outdoors and \leq 2 Mbps indoors. With this increased bandwidth availability, wireless videocommunications becomes possible. In fact, the first widely deployed mobile wireless videocommunication service was started recently in Japan [1]. This service, called Freedom of Mobile multimedia Access (FOMA), is based on the International Telecommunications Union (ITU)'s International Mobile Telephony (IMT) 2000 3G mobile communication standard. FOMA provides a 64kbps circuit-switched wireless connection for videoconferencing and a 384-kbps downlink packet-switched wireless connection for streaming video.

A concern while transmitting compressed video data over wireless channels is that the wireless channel is a noisy channel and error bursts are commonly encountered on it because of multipath fading. The effect of channel errors on compressed video data can be deleterious.

^{*} Portions reprinted, with permission, from M. Budagavi, W. R. Heinzelman, J. Webb, and R. Talluri, "Wireless MPEG-4 video communication on DSP chips," *IEEE Signal Processing Magazine*, Vol. 17, No. 1, pp. 36–53, January 2000. © 2000 IEEE.

Compressed video data are more sensitive to channel interference because of the absence of redundancy in the data. When the video bitstreams get corrupted on the wireless channel, predictive coding causes errors in the reconstructed video to propagate in time to future frames of video, and the variable-length codewords cause the decoder to easily lose synchronization with the encoder in the presence of bit errors. The end result is that the received video soon becomes unusable. Hence it becomes important that a good transport mechanism or protocol, one that provides adequate error protection to the compressed video bit stream, be used while transmitting the video data. Techniques such as forward error correction (FEC) channel coding and/or Automatic Repeat reQuest (ARQ) [2] are usually used for error protection when transporting video data over wireless channels. These techniques introduce redundancy in the transmitted data, thereby giving up some coding efficiency gains achieved by video compression. They also introduce additional delays in the system. In practice, depending on the bandwidth and delay constraints of the system, channel coding can be used to provide only a certain level of error protection to the video bit stream and it becomes necessary for the video decoder to accept some level of errors in the bit stream. Thus, it becomes essential to use error resilience techniques in the video coding scheme so that the video decoder performs satisfactorily in the presence of these errors.

Another challenge faced in wireless videocommunications is that compression and decompression of video and audio data are computationally very complex. Therefore, the processors used in the mobile phones must have high performance — they must be fast enough to play out and/or encode video in real time. Digital signal processors (DSPs) and application-specific integrated circuits (ASICs) are well suited for this task. The processors must also have low power consumption to avoid excessive battery drain and they must also be small enough to fit into compact form factors of mobile phones. Progress in microelectronics has enabled processors to satisfy these conflicting requirements. Note that the power consumption of the displays used to view the video also becomes important and it also needs to be low enough.

International standardization has also played an equally important part in facilitating wireless videocommunications. Standardization of video compression algorithms and communication systems and protocols allow devices from different manufacturers to interoperate — this brings economies of scale and mass production of equipment into picture, thereby facilitating cost-effective services.

In this article, we will focus on describing the relevant international standards that have made wireless videocommunications possible. We will cover standards that specify both the video compression as well as the *systems* for wireless videocommunications. In the next section, we start off by providing an overview of wireless videocommunication systems. We describe three categories of wireless videocommunication systems: messaging video systems, streaming video systems, and conferencing video systems. Wireless videocommunication systems are actually wireless multimedia communication systems since video is usually transmitted along with speech, audio, other multimedia data such as still pictures and documents, and control signals. In order to design a good videocommunication system, it is important to understand the interplay of video with the other components in the system. Therefore, we also provide an overview of the various components of a wireless multimedia system. The overviews provided in Section 2 will help in understanding why the various parts of wireless multimedia communication standards are required when we explain the standards in a later section. In Section 3, we talk about the Motion Pictures Experts Group (MPEG)-4 video compression standard [3]. MPEG-4 has been standardized by the International Standards Organization(ISO)/International Electrotechnical Commission (IEC). The MPEG-4 video coding standard caters to a wide range of multimedia applications covering a variety of storage media and transmission channels. We describe only those parts of the video coding standard that are suited for mobile wireless communications-this subset of the MPEG-4 video coding standard is called the Simple Profile. In Section 4, we describe systems standards specified by the Third Generation Partnership Project (3GPP) for messaging, streaming, and conferencing over 3G wireless networks. We conclude the article with discussions in Section 5.

2. OVERVIEW

There are basically three categories of wireless videocommunication systems: messaging video systems, streaming video systems, and conferencing video systems. One of the main factors that separates these three systems is the amount of playout delay in the receiver. Playout delay is the amount of time between the reception of video data in receiver and the playout of the received video data in the receiver. The playout delay determines whether a high-delay or a low-delay wireless connection is required. Another factor that separates these systems is whether both the video encoder and decoder or only the video decoder is used in the mobile phone — this determines whether a two-way or a one-way wireless video connection is required.

Figure 1 shows the block diagrams of the three wireless videocommunication systems. In messaging video systems (see Fig. 1a), a mobile phone wishing to send a video message (e.g., a vacation video clip) first creates the video message by capturing and encoding the video. It then uploads the complete video message onto a multimedia messaging service center (MMSC) along with the address of the mobile phone to which the video message is directed to. The mmsc then notifies the recipient mobile phone of the video message that has been sent to it. The recipient mobile phone then downloads the whole video message before playing it out. Video email is a variation of the messaging scheme explained above. In video email, the recipient mobile phone polls the email server to see if it has any email. If there is an email on the server, the mobile phone downloads it fully before playing it out. The playout



Figure 1. Three basic categories of wireless video systems: (a) messaging systems; (b) streaming systems; (c) conferencing systems. A solid line indicates a wireline connection and a dashed line indicates a wireless connection.

delay in the case of messaging video systems is the time taken to download the entire video message. Note that this playout delay may not be visible to the end user, since the downloading could be occurring in the background and the end user could be notified of the message only when the download is complete. Because of their download-and-play nature, it is not necessary for video messaging systems to have a low-delay connection. Also, the video encoder is not required if the mobile phone wants to have the capability of only receiving video messages.

In streaming video systems (see Fig. 1b), the video data received from the video server are buffered for a small amount of time (e.g., ~ 3 s) before being played out. This small buffer absorbs the delay jitters experienced by the video data sent on the wireless channel. The end result is that even if there is delay variation on the wireless channel, the video playout will still be smooth without any breakups. The playout delay in the case of streaming video systems is equal to the initial buffering delay. Note that streaming video systems require a connection that has a delay less than the initial buffering delay in order to have a smooth playout without any breaks. The streaming video data is sent in only one direction-from the video server to the mobile phone. In streaming video systems, a video encoder is not required on the mobile phone. The streaming video data can come from prestored video clips on the server, or they can come from live feeds of news and entertainment events. Note that the video is distributed from the source location to the video servers using a content distribution network, which is typically a wireline network.

Conferencing video systems (see Fig. 1c), which are used mainly for videoconferencing, have very strict delay requirements. The end-to-end delay must be less than 150 ms (though somewhat higher delays might still be acceptable) for the videoconference to be natural. Hence the video data that are received is played out as soon as possible. Also because of the two-way nature of the videoconference, videoconferencing systems require both the video encoder and decoder, and a two-way wireless channel for simultaneously transmitting and receiving the video data.

It is important to note that Fig. 1 illustrates wireless systems, but not wireless applications. In many cases, wireless applications can run on one or more of the wireless systems shown in Fig. 1. For example, a streaming video player *application* can be built on top of a conferencing video system or a streaming video system. The behavior of the streaming video player application will be different on the streaming and the conferencing video system. If the streaming video player application is built on top of a conferencing video system, since there is a very low-delay connection, the streamed video can be played out sooner. The uplink wireless channel (from the mobile phone to the cellular network) will remain unused in this case.

In each of three wireless video systems, video is usually transmitted along with speech/audio and other multimedia data such as images and documents. Therefore the mobile phone used for wireless videocommunications consists of various other components as shown in Fig. 2. The video codec, the audio codec, and the multimedia data blocks process (compress/decompress)

WIRELESS MPEG-4 VIDEOCOMMUNICATIONS 2975



Figure 2. Components of a general wireless multimedia phone.

the video, audio, and multimedia data used in the multimedia communication session. In addition to these blocks, we have two other important blocks: control and multiplex-demultiplex-synchronization (MDS) blocks. The control block is used to initiate and teardown the multimedia communication session. It is also used to decide the audio and video compression methods to use and the data rates to use. The MDS block is used to combine the audio, video, multimedia data, and control signals into a single stream before transmission on the wireless network. In the receiver, it is used to demultiplex the received stream to obtain the audio, video, multimedia data, and control signals which are then passed on to their respective processing blocks. The MDS block is also used to synchronize and schedule the presentation of audio, video and other multimedia data.

In the next section, we describe the video codec block and in Section 4, we will look at the various manifestations of Fig. 2 as applied to messaging, streaming, and conferencing systems standards.

3. MPEG-4 SIMPLE PROFILE VIDEO COMPRESSION¹

The Simple Profile of the MPEG-4 video standard [3] uses compression techniques similar to H.263 [5], with

 $^1\,\rm This$ section and the figures appearing in this section have been taken from Ref. 4 with some modifications



Figure 3. MPEG-4 simple profile includes error resilience tools for wireless applications. The core of MPEG-4 simple profile is the H.263 coder. Resynchronization markers, header extension code (HEC), data partitioning, and reversible VLCs provide error resilience support. From Fig. 4 of [4]. © 2000 IEEE with permission.

some additional tools for error detection and recovery. The scope of MPEG-4 simple profile is schematically shown in Fig. 3. As in H.263, video is encoded using a hybrid block motion compensation (BMC)/discrete-cosine transform (DCT) technique. Figure 4 illustrates a standard hybrid BMC/DCT video coder configuration. Pictures are coded in either *intraframe* (INTRA) or *interframe* (INTER) mode, and are called *I frames* or *P frames*, respectively. For intracoded I frames, the video image is encoded without any relation to the previous image, whereas for intercoded P frames, the current image is predicted from the previous reconstructed image using BMC, and the difference between the current image and the predicted image (referred to as the *residual image*) is encoded.



Figure 4. A standard videocoder based on block motion compensation and DCT. From Fig. 5 of [4]. © 2000 IEEE with permission

2976 WIRELESS MPEG-4 VIDEOCOMMUNICATIONS

The basic unit of information which is operated on is called a macroblock and is the data (both luminance and chrominance) corresponding to a block of 16×16 pixels. Motion information, in the form of motion vectors, is calculated for each macroblock in a P frame. MPEG-4 allows the motion vectors to have half-pixel resolution and also allows for four motion vectors per macroblock. Note that individual macroblocks within a P frame can be coded in INTRA mode. This is typically done if BMC does not give a good prediction for that macroblock. All macroblocks must also be INTRA-refreshed periodically to avoid the accumulation of numerical errors, but the INTRA refresh can be implemented asynchronously among macroblocks.

Depending on the mode of coding (INTER or INTRA) used, the macroblocks of either the image or the residual image are split into blocks of size 8×8 , which are then transformed using the DCT. The resulting DCT coefficients are quantized, run-length-encoded, and finally variablelength-coded (VLC) before transmission. Since residual image blocks often have very few nonzero quantized DCT coefficients, this method of coding achieves efficient compression. For INTER-coded macroblocks, motion information is also transmitted. Since a significant amount of correlation exists between neighboring macroblocks' motion vectors, the motion vectors are themselves predicted from already transmitted motion vectors, and the motion vector prediction error is encoded. The motion vector prediction error and the mode information are also variable-length-coded before transmission to achieve efficient compression. In the decoder, the process described above is reversed to reconstruct the video signal. Each video frame is also reconstructed in the encoder, to mimic the decoder, and to use for motion estimation of the next frame.

Because of the use of VLC, compressed video bit streams are particularly sensitive to channel errors. In VLC, the boundary between codewords is implicit. Transmission errors typically lead to an incorrect number of bits being used in VLC decoding, causing loss of synchronization with the encoder. Also, because of VLC, the location in the bit stream where the decoder detects an error is not the same as the location where the error has actually occurred. This is illustrated in Fig. 5. Once an error is detected, all the data between the resynchronization points are typically discarded. The error resilience tools in MPEG-4 simple profile basically help in minimizing the amount of data that has to be discarded whenever errors are detected. The error resilience tools included in the *simple* profile to increase the error robustness are

- Resynchronization markers
- Data partitioning
- Header extension codes (HECs)
- Reversible variable-length codes (RVLCs)

In addition to these tools, error concealment [6] should be implemented in the decoder. Also, the encoder can be implemented to limit error propagation using an adaptive INTRA refresh technique [7].

3.1. Resynchronization Markers

As mentioned earlier, a video decoder that is decoding a corrupted bit stream typically loses synchronization with the encoder due to the use of variable-length codes. MPEG-4 adopted a resynchronization strategy referred to as the "video packet" approach. Packetization allows the receiver to resynchronize with the transmitter when a burst of transmission errors corrupts too much data in an individual packet. A video packet consists of a resynchronization marker, a video packet header, and macroblock data, as shown in Fig. 6. The resynchronization marker is a unique code, consisting of a sequence of zero bits followed by a 1-bit, that cannot be emulated by the variable-length codes used in MPEG-4. Whenever an error is detected in the bit stream, the video decoder jumps to the next resynchronization marker to establish synchronization with the encoder. The video packet header contains information that helps in restarting the decoding process, such as the absolute macroblock number of the first macroblock in the video packet and the initial quantization parameter used to quantize the DCT coefficients in the packet. A third field, labeled HEC, is discussed in Section 3.4. The macroblock data part of the video packet consists of the motion vectors, DCT coefficients, and mode information for the macroblocks contained in the video packet.

marker number Quant HEC Macrobiock data

Figure 6. Resynchronization markers help in localizing the effect of errors to an MPEG-4 video packet. The header of each video packet contains all the necessary information to decode the macroblock data in the packet. From Fig. 7 of [4]. © 2000 IEEE. with permission.




The predictive encoding methods are modified so that there is no data dependency between the video packets of a frame. Each video packet can be independently decoded irrespective of whether the other video packets of the frame are received correctly. A video packet always starts at a macroblock boundary. The exact size of a video packet is not fixed by the MPEG-4 standard (the standard does specify the maximum size that a video packet can take); however, it is recommended that the size of the video packets (and hence the spacing between resynchronization markers) be approximately equal.

3.2. Data Partitioning

The data partitioning mode of MPEG-4 partitions the macroblock data within a video packet as shown in Fig. 7. For I frames, the first part contains the coding mode and six dc DCT coefficients for each macroblock (4 for luminance and 2 for chrominance) in the video packet, followed by a dc_marker (DCM) to denote the end of the first part, as shown in Fig. 7a. (Note that the zeroeth DCT coefficient is called the dc DCT coefficient, the remaining 63 DCT coefficients are called ac coefficients.) The second part contains the ac coefficients. The DCM is a 19-bit marker whose value is 110 1011 0000 0000 0001. If only the ac coefficients are lost, the dc values can be used to partially reconstruct the blocks. For P frames, the macroblock data is partitioned into a motion part and a texture part (DCT coefficients) separated by a unique motion_marker (MM), as shown in Fig. 7b. All the syntactic elements of the video packet that are required to decode motion related information are placed

(a)	Resync marker	Header	DCT DC/Mode information	DCM	Texture information
(b)	Resync marker	Header	Motion/Mode information	ММ	Texture information

Figure 7. MPEG-4 data partitioned video packet for (a) I frames and (b) P frames. Data partitioning uses additional markers (DCM and MM) and puts the most important information in the first partition of video packet, for better error concealment. From Fig. 8 of [4]. © 2000 IEEE with permission.

in the motion partition and all the remaining syntactic elements that relate to the DCT data are placed in the texture partition. The MM indicates to the decoder the end of the motion information and the beginning of the DCT information. The MM is a 17-bit marker whose value is 1 1111 0000 0000 0001. If only the texture information is lost, data partitioning allows the use of motion information to conceal errors in a more effective manner. Data partitioning thus provides a mechanism to recover more data from a corrupted video packet.

3.3. Reversible Variable-Length Codes (RVLCs)

Reversible VLCs can be used with data partitioning to recover more DCT data from a corrupted texture partition. Reversible VLCs are designed such that they can be decoded both in the forward and the backward direction. Figure 8 illustrates the steps involved in two-way decoding of RVLCs in the presence of errors. While decoding the bit stream in the forward direction, if the decoder detects an error, it can jump to the next resynchronization marker and start decoding the bit stream in the backward direction until it encounters an error. Based on the two error locations, the decoder can recover some of the data that would have otherwise been discarded. Because the error may not be detected as soon as it occurs, the decoder may conservatively discard additional bits around the corrupted region. Note that if RVLCs were not used, more data in the texture part of the video packet would have to be discarded. RVLCs thus enable the decoder to better isolate the error location in the bit stream. Note that RVLC can be used only when data partitioning is enabled.

3.4. Header Extension Code (HEC)

Important information that remains constant over a video frame, such as the spatial dimensions of the video data, the timestamps associated with the decoding and the presentation of these video data, and the type of the current frame (INTER coded/INTRA coded), are transmitted in the header at the beginning of the video frame data. If some of this information is corrupted due to channel errors, the decoder has no other recourse but to discard all the information belonging to the current



Figure 8. Reversible VLCs can be parsed in both the forward and backward directions, making it possible to recover more DCT data from a corrupted texture partition. From Fig. 9 of [4]. © 2000 IEEE with permission.

video frame. In order to reduce the sensitivity of this data, a 1-bit field called HEC is used in the video packet header. When HEC is set, the important header information that describes the video frame is repeated in the bits following the HEC. This duplicate information can be used to verify and correct the header information of the video frame. The use of HEC significantly reduces the number of discarded video frames and helps achieve a higher overall decoded video quality.

3.5. Error Concealment

The MPEG-4 standard does not specify what action the decoder should take when an error is detected. Several error concealment techniques have been developed based on temporal, spatial, or frequency-domain prediction of the lost data [6]. The simplest temporal concealment technique is macroblock copy. Under this procedure, corrupted macroblocks are replaced with collocated macroblocks from the previous frame. In practice this technique works quite satisfactorily when the amount of motion in video sequences is low, such as the head and shoulder video sequence type that arises in videoconferencing. More sophisticated temporal concealment techniques use the motion vector of the macroblock to copy the motion compensated macroblock from the previous frame. Sometimes the motion vector is available when data partitioning is used. In cases where the motion vector of the macroblock is lost, it is estimated from the motion vector of the neighboring macroblocks. However, temporal concealment cannot be used for the first frame (which is an I frame), and may yield poor results for intracoded macroblocks or areas of high motion. In such cases spatial domain error concealment techniques, wherein lost blocks are interpolated from correctly received neighboring blocks in the video frame, have to be used. Concealment in the spatial domain typically involves more computation because of the use of pixel-domain interpolation. In some cases, frequency-domain interpolation may be more convenient, by estimating the dc value and possibly some low-order ac DCT coefficients.

3.6. Adaptive INTRA Refresh (AIR)

AIR is a standard-compatible encoder technique for limiting error propagation by using nonpredictive INTRA coding [7]. INTRA refresh forcefully encodes some macroblocks in INTRA mode to flush out possible errors. INTRA refresh is very effective in stopping the propagation of errors, but it comes at the cost of a large overhead; coding a macroblock in INTRA mode typically requires many more bits than coding in INTER mode. Hence the INTRA refresh technique has to be used judiciously.

AIR adaptively performs INTRA refresh based on the motion in the scene. For areas with low motion, simple temporal error concealment works quite effectively. Since the high motion areas can propagate errors to many macroblocks, any persistent error in the high motion area becomes very noticeable. The AIR technique of MPEG-4 INTRA refreshes the motion areas more frequently, thereby allowing the possibly corrupted high motion areas to recover quickly from errors.

4. WIRELESS VIDEOCOMMUNICATION SYSTEM STANDARDS

In this section we briefly describe the wireless videocommunication system standards recommended by 3GPP for messaging, streaming, and conferencing. The standards are Multimedia Messaging Services (MMS) standard [8] for messaging, the Real-time Streaming Protocol (RTSP) standard [9,10] for streaming, the Session Initiation Protocol (SIP) standard [11,12] for videoconferencing over packet-switched networks, and the 3G-324 standard [13] for videoconferencing over circuit-switched networks.

It is useful to understand the characteristics of the network types used for these standards first before reading about the standards. The MMS, RTSP, and SIP standards are used over packet-switched networks whereas the 3G-324 standard is used over circuitswitched networks. Circuit-switched networks allocate a dedicated amount of bandwidth to the connection and hence they provide a predictable-delay connection. On the other hand, on packet-switched networks, data is packetized and transmitted over shared bandwidth and thus a predictable timing of data delivery cannot be guaranteed. The types of channel impairments observed on these two types of networks are different. On circuitswitched networks, the transmission errors experienced are in the form of bit errors, whereas on packet-switched networks, the transmission errors experienced are in the form of packet losses. On packet-switched networks, the predominantly used network layer protocol is the Internet Protocol (IP). There are two transport layer protocols developed for use with IP: the Transmission Control Protocol (TCP) and the User Datagram Protocol (UDP). TCP provides a reliable point-to-point service for delivery of packet information in proper sequence, whereas UDP simply provides a service for delivering packets to the destination without guarantee. TCP uses retransmission of lost packets to guarantee delivery. TCP is found to be inappropriate for real-time transport of audio video information because retransmissions may result in indeterminate delays leading to discernible distortions and gaps in the real-time playout of the audio/videostreams. In contrast, UDP does not have the problem of indeterminate delays because it does not use retransmission. However, the problem in using UDP for transmitting media is that UDP does not provide sequence numbers to transmitted packets. Hence UDP packets can get delivered out-of-order and packet loss might go undetected.

Before proceeding ahead in this section, it is also useful to revisit Fig. 2 and Section 2. All the standards follow the overall architecture of Fig. 2. They all differ in the type of control and multiplex-demultiplex-synchronization blocks they use.

4.1. Multimedia Messaging Services (MMS) Standard

Figure 9 shows the MMS [8] protocol stack for multimedia messaging over wireless IP networks. Each MMS multimedia message consists of a MMS header and an optional MMS body. The MMS header is used for signaling information between the mobile phone and the multimedia messaging service center (MMSC), and the MMS body



Figure 9. MMS protocol stack for multimedia messaging.

is used to carry the actual multimedia message data. The MMS header includes information on the type of the MMS message, specifically, whether it is a request from the mobile phone to the MMSC, a notification from the MMSC to the mobile phone, or a confirmation response to a request/notification. The header also contains the destination address, the date when the multimedia message was created, the address of the originating mobile phone, the version of the MMS protocol, and more such fields. The MMS body contains the media data and also the layout information, that is, information on where the various media components should be displayed on the screen and also as to when they should be played out and in what order. The presentation layout is specified using the Synchronized Multimedia Integration Language (SMIL) [14]. If audio and video must be played out in a synchronized fashion, then they must be packaged together using the MPEG-4 file format [15] first. The multiple media elements and the SMIL description are combined into a single composite entity using the Multipurpose Internet Mail Extensions (MIME) multipart format [16]. This final MIME encapsulated data forms the MMS body. The whole MMS message (MMS headers and the MMS body) is transmitted using transport protocols used for emails. The email transport protocols are layered on TCP which provides a reliable connection. TCP can be used in messaging systems because messaging systems can tolerate delays. At the receiving end, after the entire MMS message has been received, the mobile phone extracts the multimedia data and the SMIL description from the MIME encapsulated MMS body, and plays out the media according to the presentation information in the SMIL description.

Comparing Fig. 9 to Fig. 2 at a high level, the MMS headers forms the control block and the combination of SMIL, MPEG-4 file format, and MIME forms the multiplex-demultiplex-synchronization block.

4.2. Real-Time Streaming Protocol (RTSP)

The protocol stack for a RTSP-based [9,10] streaming media player is shown in Fig. 10. RTSP specifies a textbased protocol for exchanging control information with the video server. Control messages are used to establish the



Figure 10. RTSP protocol stack for streaming video.

streaming session and to signal the type and format of the media to be used in the streaming session. They are also used for functions such as pausing, fast forwarding, and stopping the media playout. At the beginning of the streaming session, the mobile first sends a RTSP message to the video server specifying the media clip that it wants to watch. This is similar to sending a Webpage address to the Webserver for downloading the Webpage from the Webserver. The video server replies back, providing information on the type and format of the media in the media clip. The Session Description Protocol (SDP) [17] is used to provide this information. The SDP information is enclosed in the response from the server. The mobile phone looks at the enclosed SDP message and decides if it is able to decode the types of media present in the media clip. If it can, it then proceeds and issues a RTSP request to the video server to start streaming the media clip. The video server then starts streaming the media to the mobile phone. The mobile phone typically buffers the media for a short duration of time (e.g., 3 s) before playing them out. At anytime the streaming has to be stopped or paused, the mobile phone sends a RTSP request informing the video server to do so. The RTSP messages are usually transmitted reliably by using TCP.

The media are usually transmitted using UDP for prompt delivery. UDP does not provide timestamp information that is required in the playout of the media. Also as was stated earlier, by using UDP, media packets can get lost and can be delivered out-of-order. Hence the Real-time Transport Protocol (RTP) [18] is used on top of UDP to overcome the shortcomings of UDP. The RTP packet headers contain timestamps and sequence number information. The sequence number enables detection of packet loss and out-of-order packet delivery, and the timestamp provides timing information required in the playout of media.

Comparing Fig. 10 to Fig. 2 at a high level, the combination of RTSP and SDP forms the control block. Synchronization is carried out by RTP. There is no explicit multiplex–demultiplex block since the multiplexing and demultiplexing is carried out at the network level below IP.

4.3. Session Initiation Protocol (SIP)

Figure 11 shows the protocol stack for a SIP-based [11,12] mobile phone for videoconferencing over wireless IP



Figure 11. SIP protocol stack for two-way videoconferencing over packet-switched networks.

networks. As in the case of the RTSP streaming media player, RTP is used for transporting media. The SIP videoconferencing terminal contains both the encoder and the decoder for audio/video, unlike the RTSP streaming media player, which contains only the audio/video decoders.

The SIP protocol is used for signaling call setup and teardown. The SIP protocol can be layered on top of TCP or UDP. SIP also uses textual encoding of control messages. A SIP-based mobile phone wishing to place a call sends a message to the remote end inviting it to a call. Along with the invite message, the mobile phone also sends a SDP message describing the types and formats of media that it can receive and transmit during the call. If the remote end is ready to accept the call, it sends an acknowledgment to the invite. In the acknowledgment, the remote end sends back a SDP message indicating the media types and formats it can receive and transmit. Using the SDP messages, both endpoints can then decide on the media format that each will use for transmission and then start transmitting the media using RTP. Either end can end the call by sending a "Bye" message.

Comparing Fig. 11 to Fig. 2 at a high level, the combination of SIP and SDP forms the control block. Synchronization is carried out by RTP. There is no explicit multiplex-demultiplex block since the multiplexing and demultiplexing is carried out at the network level below IP.

4.4. 3G.324

For videoconferencing over circuit-switched wireless networks, 3GPP specifies the use of the 3G-324 standard [13]. The 3G-324 standard is based on the ITU standard for circuit-switched multimedia communications—H.324 [19]. The protocol stack for a 3G-324 videoconferencing terminal is shown in Fig. 12. Video, audio,



Figure 12. 3G.324 protocol stack for two-way videoconferencing over circuit-switched networks.

and the control information are sent on distinct logical channels. The H.245 standard [20] is used to negotiate the media to be used in the call and also to set up the logical channels for the media. The H.223 standard [21] determines the way in which the logical channels are mixed into a single bit stream before transmission over the wireless channel. The H.223 standard was originally designed for operation over the benign public switched telephone networks. H.223 was later extended for operation over error-prone wireless channels. H.223 consists of two layers: the adaptation layer and multiplex layer. The adaptation layer allows for the use of both FEC and ARQ to protect the media being transmitted. This error protection is in addition to what might be provided by the wireless network. The multiplex layer is responsible for multiplexing the various logical channels into a single bit stream.

Comparing Fig. 12 to Fig. 2 at a high level, H.245 forms the control block and H.223 forms the multiplex-demultiplex-synchronization block.

5. DISCUSSION

In this article we provided an overview of wireless videocommunications and described the various international standards that have made it possible. In addition to video compression, error resilience techniques for video and efficient mechanisms for video transport are important in wireless videocommunications. We discussed the Simple Profile of MPEG-4 video coding standard which introduces several error resilience tools aimed at containing the effect of transmission errors in the video bit stream. In practice, video is usually transmitted along with speech, audio, multimedia data such as images and documents, and control signals to form a complete multimedia communication system. It is important to understand the interplay of video with the other components of the multimedia communication system. So we also provided an overview of the following systems standards that have been recommended by 3GPP for use over third generation wireless networks: MMS for multimedia messaging, RTSP for streaming, SIP for videoconferencing over packet-switched wireless networks, and 3G.324 for videoconferencing over circuitswitched wireless networks. It should be noted that though these standards were specifically described for use on wireless networks, many of them can be used and are being used on wireline networks too.

In addition to the standards described in this article, there are other video coding standards and proprietary techniques that can/will be used on wireless networks. ITU has also extended the H.263 to provide support for error resilience. In 3GPP standards, baseline H.263 is the mandatory video coder that has to be supported. Both MPEG-4 and the extensions to H.263 (called H.263++) are optional. The first commercial wireless video deployment—FOMA [1]—uses MPEG-4. For wireless streaming applications, proprietary video compression standards such as Quicktime, RealVideo and Microsoft Windows Media Video might also be used because of the wide availability of existing content in those formats on the World Wide Web.

Wireless videocommunications is a relatively new field, and there is a lot of ongoing research activity for improving the overall video quality. Techniques such as joint-source channel coding, where the amount of bits allocated to source coding and channel coding are adaptively varied according to channel conditions, and unequal error protection (UEP), where the level of error protection is varied per the importance of the data, are being studied. Layered video coding, where the video is split into a base layer and several enhancement layers that provide incremental quality improvement over the layers below them, can be used in conjunction with UEP with the base layer being protected the most. Low-complexity (and hence low-power) video compression is another important field of research. In addition to research in video compression, research activities in the fields of low-power semiconductors and displays, wireless communications and networking, are all simultaneously enabling wireless video to become a compelling application on mobile phones.

BIOGRAPHY

Madhukar Budagavi received the B.E. degree (first class with distinction) in electronics and communications engineering from the Regional Engineering College, Trichy, India, in 1991, and the M.Sc.(Engg.) degree in electrical engineering from Indian Institute of Science, Bangalore, India, in 1993, and the Ph.D. degree in electrical engineering from Texas A&M University, College station, Texas (USA), in 1998.

From 1993 to 1995, he was first a Software Engineer and then a Senior Software Engineer in Motorola India Electronics Ltd., primarily developing DSP software and algorithms for the Motorola DSP chips. Since 1998, he has been a Member of Technical Staff with the Texas Instruments DSP Solutions R&D center, working on MPEG-4 and protocols for wireless videocommunications. His research interests include video coding, speech coding, and wireless and Internet multimedia communications.

BIBLIOGRAPHY

- NTT DoCoMo, Freedom of Mobile Multimedia Access (online), NTT DoCoMo (2002); http://foma.nttdocomo.co.jp/english/ catalog/network/index.html (March 15, 2002).
- S. Lin and D. J. Costello Jr., Error control coding: Fundamentals and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1983.
- 3. International Standards Organization(ISO)/International Electrotechnical Commission (IEC), Information Technology—Coding of Audio-Visual Objects—Part 2: Visual, ISO/IEC 14496-2, 1999.
- M. Budagavi, W. Rabiner Heinzelman, J. Webb, and R. Talluri, Wireless MPEG-4 video communication on DSP chips, *IEEE Signal Process. Mag.* 17: 36-53 (Jan. 2000).
- International Telecommunications Union (ITU), Video Coding for Low Bitrate Communication, Recommendation H.263, 1996.
- Y. Wang and Q.-F. Zhu, Error control and concealment for video communications: A review, *Proc. IEEE* 86: 974–997 (May 1998).

- K. Imura and Y. Machida, Error resilient video coding schemes for real-time and low-bitrate mobile communications, *Signal Process. Image Commun.* 14: 519-530 (May 1999).
- 3rd Generation Partnership Project, Multimedia Messaging Services (MMS): *Functional Description*, Technical Specification 23.140, V5.1.0, 2001.
- 3rd Generation Partnership Project, Transparent End-to-End Packet Switched Streaming Service (PSS): Protocols and Codecs, Technical Specification 26.234, V4.2.0, 2001.
- 10. Internet Engineering Task Force, *Real Time Streaming Protocol* (*RTSP*), RFC 2326.
- 3rd Generation Partnership Project, Packet Switched Conversational Multimedia Applications: Default Codecs, Technical Specification 26.235, V5.0.0, 2001.
- 12. Internet Engineering Task Force, SIP, Session Initiation Protocol, RFC 2543.
- 3rd Generation Partnership Project, Codec for Circuit Switched Multimedia Telephony Service: Modification to H.324, Technical Specification 26.111, V3.4.0, 2000.
- W3C, Synchronized Multimedia Integration Language (SMIL 2.0), http://www.w3.org/TR/2001/REC-smil20-20010807/ (Aug. 2001).
- 15. International Standards Organization(ISO)/International Electrotechnical Commission (IEC), Information Technology—Coding of Audio-visual Objects—Part 1: Systems, ISO/IEC 14496-1, 2001.
- 16. Internet Engineering Task Force, Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types, RFC 2046.
- 17. Internet Engineering Task Force, SDP, Session Description Protocol, RFC 2327.
- Internet Engineering Task Force, RTP, A Transport Protocol for Real-Time Applications, RFC 1889.
- International Telecommunications Union—Telecommunications Standardization Sector, Terminal for Low Bit Rate Multimedia Communications, Recommendation H.324, 1998.
- International Telecommunications Union—Telecommunications Standardization Sector, Control Protocol for Multimedia Communication, Recommendation H.245, 1999.
- 21. International Telecommunications Union—Telecommunications Standardization Sector, *Multiplexing Protocol for Low Bitrate Multimedia Communication*, Recommendation H.223 and Annex A, B, C, 1997.

WIRELESS PACKET DATA

KRISHNA BALACHANDRAN KENNETH BUDKA WEI LUO Lucent Technologies Bells Laboratories Holmdel, New Jersey

1. INTRODUCTION

Wireless packet data is defined as the transfer of information over wireless links that are shared dynamically by multiple users without dedicating wireless resources to individual users during periods of inactivity.

In traditional circuit-mode cellular voice services, radio resources (e.g., frequencies, time slots, or codes) are assigned exclusively to a single user for the duration of a call. As a result, resources remain assigned during periods of inactivity. Circuit mode voice service is fairly wasteful of network resources. Studies of conversational patterns, for example, indicate that the amount of time speakers actually speak accounts for only 40-50% of the duration of the call. The remaining time is consumed by pauses in conversation. Due to the bursty nature of data transmissions, periods of activity for data services users tend to be much lower (e.g., 10-15%). Under these traffic assumptions, packet-mode operation allows much more efficient utilization of radio resources than circuit mode. In packet mode, a stream of bits is broken down into smaller units, or packets, which are individually transferred through the network. The communication resource may be dynamically shared by multiple users at any given time. Circuit mode operation wastes resources during periods of inactivity. With packet mode operation, the communication resource is shared by multiple users, and an active user can take advantage of the inactive periods of other users.

The most commonly used electromagnetic wireless access media include radio frequency (RF) waves and lightwaves. When the frequency is below 1 THz (terahertz), the electromagnetic waves are referred to as RF, and above 1 THz, they are called lightwaves. The most commonly used RF band for wireless communications is from 800 MHz to 100 GHz. The most commonly used lightwave wavelengths for wireless communications are infrared, with wavelengths ranging from 870 to 900 nm. The range of RF spectrum less than 800 MHz has been used by radio, TV broadcast, and transportation. Generally speaking, the higher the frequency, the more directional the propagation of electromagnetic waves and the shorter the range of the signal. At one extreme, lightwaves are usually limited to short range line-of-sight communications, such as indoor wireless and point-to-point inter-building communications. At the other extreme, RF spectrum at about 1 or 2 GHz is widely used for cellular wireless communications over large coverage areas because this spectrum does not experience interference from other services and it also provides good coverage in both rural and urban areas without any line-of-sight requirements. As a result, multiple users at different locations can easily share the radio channel. This is well suited for point-to-multipoint, broadcast or multicast communication.

The transmission characteristics of wireless channels are random, time-varying, and difficult to predict. Consequently, the data rate that can be achieved varies depending on the prevailing channel quality. Therefore, it is important to employ link adaptation algorithms that can measure the channel and quickly adapt to the physicallayer parameters in order to maximize the throughput under the prevailing channel conditions. Furthermore, wireless packet data channels are shared by multiple users, and the user throughput performance or system capacity is strongly dependent on the resource allocation and scheduling schemes employed. In general, resource allocation, scheduling, and link adaptation schemes should be carefully designed in order to maximize the user throughput and/or the system capacity.

Portability and mobility are two major advantages of wireless packet data. Wireless local area networks (WLANs), for example, IEEE 802.11 [1], were originally conceived to serve the purpose of portability within the office. Most lightwave-based wireless packet data services are also used to avoid wiring. Unlike WLANs, cellular packet data systems typically support lower data rates but provide wide-area coverage and support mobility. Third Generation (3G) networks will also be able to coarsely track user locations and provide location based services. Cellular packet data networks have evolved from cellular voice networks, and have inherited mobility support from voice networks. WLANs, however, were conceived in order to provide Ethernet-like connectivity without wires and have focused more on providing wireless access to the wired network. The support of mobility is a task left to Internet Engineering Task Force (IETF) protocols such as mobile IP. The distinction between cellular packet data networks and WLANs may blur and finally disappear as the convergence of the two types of networks becomes more evident.

The main purpose of wireless packet data is to connect mobile users to the fixed network and to connect mobile users to each other. In the wired network, packets generated by end users are directly forwarded to the destination if the destination is within the same local area network (LAN). Otherwise, packets are forwarded through a predetermined route or are routed hop by hop to the final destination. Wireless access usually refers to the "last mile" connection of the end-user devices to the network. Wireless access can be provided using a traditional cellular approach or alternatively, using wireless LANs. The cellular approach assumes that a mobile terminal communicates with a base station over a radio link. The base stations are connected to each other and to the Internet through a wired network. The wireless LAN approach assumes a more ad hoc structure, in which terminals may communicate with each other directly without the help of a base station or wired network. However, if a wireless terminal needs access to the Internet, the wireless LAN should provide an access point that is connected to the wired network. The cellular network model and ad hoc wireless LAN model are illustrated in Fig. 1.

2. GENERIC WIRELESS PACKET DATA SYSTEM ARCHITECTURE

Figure 2 shows a generic wireless packet data system protocol stack. Layer 1 represents the physical layer, which is responsible for radio signal transmission and reception. Layer 2 includes the medium access control (MAC) and radio link control (RLC) layers. The MAC layer allocates airlink resources to mobiles, thus allowing them to share the wireless link. The RLC protocol provides reliable in-sequence packet delivery when configured to operate in an acknowledged mode. Layer 3 provides



Figure 1. Two basic network connection structures. In cellular networks, mobile stations (MSs) are connected to the base station (BS) via a wireless link. The BS is connected to a gateway that controls the data transfer between the radio access network and the rest of the wired network. In ad hoc networks, MSs are connected to each other directly through radio or infrared links. A terminal can route data to other terminals with which radio links can be established so that terminals that are a few hops away can communicate with each other.

Figure 2. Protocol stack for wireless packet data (signaling plane protocols are shaded).

functions that connect radio access networks to the wired network, such as radio resource control, mobility management and routing within and beyond the radio access networks. In addition, layer 3 may also perform header and/or payload adaptation. The radius adaptation or convergence sublayer is also referred to as the logical link central (LLC) layer. Ciphering may be performed in layer 2 or in layer 3. Layer 4, the transport layer, provides reliable end-to-end data transfer and end-to-end flow control, if desired. The application layer denotes end user applications (e.g., HTTP for Web browsing).

All four layers of the protocol stack are implemented in mobile stations. On the fixed network side, however, portions of the protocol stack are implemented on different network elements. The choice of how to partition the protocol stack is influenced by technology requirements and/or constraints imposed by legacy systems. In General Packet Radio Service (GPRS) and Universal Mobile Telecommunications System (UMTS) systems, for example (Fig. 3) [2], the base station is responsible for physical-layer functions such as channel coding/decoding, modulation/demodulation, pulseshaping and transmission; MAC, RLC, and some radio resource control functions are in a remote radio network controller, which controls several base stations. The Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN) are responsible for session management (i.e., packet data service and context activation, authentication, and charging), mobility management and routing. Any two mobile stations that are served by the same GGSN can communicate with each other directly through the wired backbone (i.e., network of SGSNs) of the wireless packet data network. In a WLAN configuration, the base station (or access point) includes all user plane protocols up to layer 3. There is no dedicated wired backbone network in this case. Instead, the access point acts like any IP router within the wired network, and IP-based protocols specified by the IETF are used for mobility and session management (mobile IP, AAA (Authentication, Authorization and Accounting) protocols, etc.).

Ideally, wireless packet data systems should provide at least the same services that are currently provided over



Figure 3. UMTS system architecture. A radio network controller (RNC) controls several base stations. Several RNCs are connected to a Serving GPRS Support Node (SGSN). The wired network of SGSNs is referred to as the "backbone" or "core network." The Gateway GPRS Support Node (GGSN) serves as a gateway between the wired backbone and the Internet or other virtual private networks.

wired networks. In addition, there may be services that are specifically targeted at mobile users. From an application perspective, wireless packet data systems should be able to provide these services with acceptable quality. From a business perspective, it is desirable to provide these services to as many users as possible. These objectives need to be satisfied under constraints on spectrum availability, cost and complexity. Furthermore, services targeted at mobile terminals should be enabled through small and inexpensive terminals with long battery life. These issues pose several challenging problems for wireless packet data system design, as will be discussed later.

3. QUALITY OF SERVICE REQUIREMENTS

Wireless packet data technologies support data transfer services with a wide range of quality of service (QoS) levels. End-to-end QoS requirements (e.g., loss and delay) can be broken down into corresponding requirements on the wireless data network (i.e., user equipment to a wireless gateway node) and on the fixed network. The QoS classes and attributes employed by the wireless data network should span a wide range of possible applications. Because of the limited amount of bandwidth available at the air interface, support of QoS must not involve significant overhead or complexity and should allow efficient utilization of resources.

The third-generation partnership project (3GPP) specifications define four QoS classes [3]:

- Conversational Class. The requirements for this class are similar to those for conventional telephony. In order to maintain acceptable quality, there are strict limits on transfer delay and loss rate (1-3%). Furthermore, the time relation between information entities (e.g., speech frames) needs to be preserved.
- Streaming Class. Streaming audio and video applications fall into this category. Streaming applications are more tolerant of packet transfer delays than conversational applications. The time relations (variation) between information entities (i.e., samples, packets) within a flow are to be preserved. In addition, loss rates must be low (1-3%).
- *Interactive Class.* Web browsing, database queries and other applications that follow a request/response pattern are considered interactive. Payload content must be preserved (i.e., lossless transfer). In addition, limits are placed on acceptable round-trip delay
- *Background Class.* This class of traffic is delay insensitive (i.e., best effort). Examples include electronic mail (email) and short message service (SMS). This class requires that the payload content be preserved but does not have stringent delay requirements.

Because of signaling and packet transfer delays, airlink error detection and recovery takes time. The stringent delay requirements for voice services, however, do not allow sufficient time for error recovery. Due to advances in speech coding, the rates demanded by voice services are quite low (<12-16 kbps). The typical approach is to pad transmissions with enough redundancy to allow operation under poor channel conditions. The channel coding is fixed, and no retransmissions are allowed. This error avoidance approach is not well suited to the support of data services. Since the delay requirements for data services are typically more relaxed than voice, error detection and recovery techniques can be used. It is inefficient to use a fixed amount of redundancy independent of the actual delay requirements and the prevailing channel quality. Higher data rates can be supported if error recovery is carried out using an automatic repeat request (ARQ) scheme.

For best-effort data services, the radio link control (RLC) layer typically allows *full recovery* [i.e., service data units (SDUs) are not discarded and there is no limit on the maximum number of retransmissions] and delivers data in sequence to the higher layer. This scheme is best suited to the support of best-effort data services over wireless links, and not to applications such as streaming where the time relation between information entities needs to be preserved. For streaming, a selective ARQ scheme with SDU discard and limited retransmission (i.e., partial recovery) capability can achieve better performance.

4. WIRELESS PACKET DATA TECHNOLOGIES

4.1. Performance Measures

4.1.1. Delay. From the user perspective, it is meaningful to consider measures such as the transfer delay for a speech frame, file, web object or webpage. The air interface is often the largest contributor to "end-to-end delay" perceived at the application layer. A widely used measure of performance for protocol performance in wireless data systems is the SDU delivery delay over the air interface (i.e., delay between peer RLC protocol layers).

The RLC SDU delivery delay may be defined as the time between SDU arrival at the transmitter RLC to insequence delivery by the receiver RLC. The RLC SDU delivery serves as a good measure of performance over the air interface for streaming, interactive, and background data services.

For data traffic with varying SDU sizes, a more appropriate measure of performance is the ratio of the SDU delivery delay to SDU size. This is typically known as the *normalized* SDU delivery delay.

SDU delay is a random quantity. Delay jitter (variance) is important when constant delivery rate is required. If the transmission control protocol (TCP) is used for end-to-end error detection and recovery, for example, the delay jitter must kept as small as possible. This will be elaborated on later. Another metric is delay percentiles, specifying the percentage of transferred packets that experience a delay exceeding a certain value. This is often used in the case when a packet delay exceeds a certain threshold and the packet is dropped.

4.1.2. Throughput. For a given choice of physical-layer parameters (coding, spreading, modulation), an upper bound on the expected user throughput over the air interface is given by $R \cdot (1 - b)$, where *R* denotes the peak

data rate and *b* denotes the expected error rate of each RLC packet. Throughput can be improved further when selective hybrid ARQ (i.e., incremental redundancy, which will be discussed later) is employed for radio link control.

There are usually two types of throughput measures mentioned in the literature. One is user-perceived throughput, which is usually used as one of the metrics to quantify the quality of service provided to the mobile user. The user-perceived throughput is the individual user's data throughput when the user has data to transmit. Wireless packet data channels are shared by multiple users and the user-perceived throughput computation should additionally account for the queuing time. In the literature, the data rate quoted for the system usually refers to user-perceived throughput. The other type of throughput is aggregated throughput or system throughput, which is usually used to quantify the system capacity or system load. The aggregated throughput is the amount of bits successfully transmitted per unit time in the whole system that includes all users. User-perceived throughput usually should be used together with system aggregated throughput, because in many cases, system aggregated throughput is proportional to the system loading and the userperceived throughput decreases when the system loading increases.

4.1.3. Packet Loss and In-Sequence Delivery. Wireless packet data systems typically require in-sequence SDU delivery over the air interface. If selective ARQ is employed, this translates into a requirement on the RLC protocol. Note that in-sequence delivery does not imply lossless data transfer.

SDU loss can occur in one of the following ways:

- Buffer overflow at the transmitter.
- SDU discard at the transmitter if the delivery delay requirements cannot be met.

The loss rate requirement depends on the QoS requirements of the application.

4.1.4. Coverage and Mobility Support. It is desirable to satisfy a minimum throughput or a maximum SDU delivery delay (or normalized SDU delivery delay) for a large fraction (90-95%) of users in the network.

Most cellular data networks allow users to roam over a wide area and attempt to provide an "always on" experience for the end user (i.e., mobility management is handled without significant impact on the service). In addition, for real-time services, handovers should occur seamlessly as a mobile terminal moves from the coverage area of one sector to another. This is achieved by imposing very stringent delay requirements on a handover. Generally speaking, there is a tradeoff between providing high-throughput service and providing large coverage and good mobility support. For example, at the time when this article is written, WLANs provide a data rate as high as 11 Mbps (with future standards targeting rates up to 54 Mbps), but coverage is typically limited to indoor environments with no mobility. On the contrary, the 3G wireless networks can support only 384 kbps, optimistically, but with almost seamless coverage and mobility support.

4.1.5. Security. Wireless packet data systems must provide safeguards against unauthorized network access and protect users' privacy. This is achieved through security protocols that carry out authentication and ciphering of user data and protect the integrity of control information.

To prevent unauthorized access, users need to be authenticated before they can access the wireless network. This typically involves comparison of a user-provided unique equipment ID number with ID numbers stored and maintained by the network. The network may also ask for a password to check the user's identity.

The content of the user's data, location, identity, and data usage pattern must all be protected. The use of wireless links makes it easier for casual eavesdroppers to infringe on a user's privacy. The objective of the security provided through wireless packet data is to provide security at levels comparable to wireline data service. This requires that user data be encrypted before they are sent over the air. In addition, the network should try to reduce the times that the mobile user sends its unique ID number over the air. This prevents eavesdroppers from deriving a user's location and data usage pattern. Once a user is registered with the network, the network can assign a temporary ID for the user and the temporary ID number can be used for authentication. It should be emphasized that the wireless data networks do not provide end-to-end security. Such security must be provided by the users/applications themselves or IP-based security protocols (e.g., IPSec).

4.1.6. Energy Efficiency. Many mobile terminals are powered by batteries and energy efficiency is quite important. Low mobile terminal power consumption is critical in order to lengthen communication time and to prevent mobile devices from overheating.

The definition of energy efficiency is not as simple as it appears to be. At first glance, energy efficiency can be defined as the amount of energy expended per information bit received. But this definition does not capture the data rate. Generally speaking, the higher the data rate, the higher the amount of energy expended per unit of information bits received. This is due mainly to two effects: (1) the higher data rate implies higher receiver processing requirements and (2) increasing data rate over the air interface requires higher transmission power per bit. According to Shannon theory, the required transmission power is an exponentially increasing function of the supported data rate. If the wireless link is shared by multiple users in certain ways, increasing one user's data rate leads to a higher interference level being experienced by other users. Therefore, the other users need more transmission power to combat the increased interference level. From a system perspective, if multiuser interference is a concern, the higher the aggregated throughput of the system, the higher the energy requirement per information bit. A

good indication of energy efficiency, therefore, is a power consumption function corresponding to the supported data rate.

Reduction of power consumption in wireless packet data involves optimization of all aspects of the system and circuit design, from hardware and software to communication protocols. For example, on the RF portion of a mobile terminal's circuitry, the use of low peak-toaverage modulation schemes and nonlinear amplifiers improve amplifier efficiency and saves energy. On the digital logic part of a mobile terminal's circuitry, lowvoltage and low-clock-frequency circuits are preferred for power reduction. Integration of the system into a small number of chips is desirable to reduce I/O power consumption. Discontinuous transmission and reception¹ modes that allow the mobile to go idle if there are no data to send or receive are very useful in reducing power consumption. All the power reduction approaches discussed must ensure that they do not compromise performance.

4.2. Technology Methods

4.2.1. Link Adaptation and Incremental Redundancy. Link adaptation is a technique that uses channel quality measurements in order to select the physical-layer parameters such as modulation, coding, and spreading that are needed in order to achieve the highest throughput under delay constraints [4]. The RLC block sizes and physical-layer parameters should be chosen in such a way that blocks can be retransmitted without significant overhead even if the physical-layer parameters used for the initial transmission are different from those used for retransmission.

Incremental redundancy is a technique that can be applied in conjunction with link adaptation in order to achieve higher throughput under a wide range of operating conditions. For each set of physical-layer parameters, a set of puncturing schemes (P1, P2, P3, etc.) achieving the same code rate are defined. All puncturing is carried out on the same rate 1/N "mother code." The initial transmission of a block consists of the bits obtained by applying the puncturing scheme P1 (for the chosen physical-layer parameters) to the rate 1/Nencoded data. On receiving a negative acknowledgment

¹ Discontinuous reception is commonly referred to as "sleep mode."

for the RLC block, additional coded bits (i.e., the output of the rate 1/N encoded data punctured with scheme, P2, corresponding to the prevailing set of physical-layer parameters) are transmitted. If all the punctured versions of the encoded data block have been transmitted, the cycle is repeated, starting again with P1. If the receiver does not have sufficient memory for incremental redundancy operation, it can attempt to decode the data by using the information corresponding to just P1, or P2, or P3. This corresponds to the pure link adaptation case. For incremental redundancy operation, the receiver must have sufficient memory in order to store soft information corresponding to RLC blocks that have not yet been decoded successfully. Each time the receiver obtains additional coded bits, it attempts soft-decision decoding using this redundant information in addition to previously stored soft information corresponding to the same RLC block(s).

The individual puncturing schemes are designed to be as disjunctive (or nonoverlapping) as possible in order to achieve good performance with incremental redundancy. In addition, to achieve good performance with pure link adaptation, the puncturing schemes are designed to ensure that the individual schemes (P1, P2, P3) achieve comparable error performance.

4.2.2. Peak Picking Scheduling. In a cellular environment, different users sharing the same airlink will observe different link performance. In addition, the performance observed by each user may vary in time due to changes in interference levels and shadow fading.

Figure 4 shows an idealized plot of logical link control (LLC)-layer efficiency variations as a function of time as link adaptation tracks the changes in link quality.

To increase overall system capacity, a scheduler can dynamically allocate larger portions of airlink resources to those users with high link efficiencies, a technique known as "peak picking." To help understand the role peak picking plays in the design of airlink schedulers, it is helpful to look at two extremes:

• One naive scheduling algorithm is to allocate all network resources to the mobile with the highest link efficiency. Such an algorithm maximizes the total amount of data carried over the airlink. However, such an algorithm is woefully impractical. One problem with such a scheduler is fairness. Using

Figure 4. Peak picking can substantially improve overall system capacity. The curves represent the throughputs that mobile 1 and mobile 2 would receive if they were given all airlink blocks. With peak picking, network operators will be able to generate more revenue from the airlink.



such a scheme, it is possible that one user in the cell is granted all airlink bandwidth, while all other users starve. Such bandwidth starvation can have disastrous effects on the performance of higher-layer protocols, such as TCP. Such starvation can result in frequent TCP retransmissions and TCP connection failures: focusing on maximizing airlink throughput alone may end up causing transmission inefficiency at higher layers. In addition, it is often necessary for the network to periodically schedule transmissions to/from a mobile so that it has current information needed for power control and link adaptation to function properly.

• At the other extreme is simple round-robin scheduling, in which each user is given an equal fraction of airlink blocks. Such a scheduling scheme does not take advantage of peak picking.

Effective schedulers employing peak picking lie somewhere between these two extremes. The design of such schedulers is currently an active area of research.

4.2.3. Header Compression. Wireless data networking technologies are designed to make most efficient use of airlink spectrum. Link adaptation and incremental redundancy techniques are one way of doing this. Another way of increasing link efficiency is to shrink the size of user data packets carried over the wireless data network's airlink. Two additional tools are used for this purpose: packet header compression and packet payload compression.

Each TCP/IP packet contains a 20-byte TCP header and a 20-byte IP header. For small payloads, 40 bytes of overhead is a large price to pay. For this reason, wireless data networks typically employ techniques to compress the headers of each network-layer packet carried over the airlink. Such schemes are based on the packet header compression scheme developed by Van Jacobsen [5]. Taking advantage of the fact that many of the fields in TCP/IP packet change little over the lifetime of a TCP connection. Van Jacobsen's header compression algorithm can reduce the amount of data needed to carry a TCP/IP packet header from 40 to 3 bytes. Similar techniques are also being proposed to reduce the headers used by UDP/IP headers to enable efficient carrying of audio streaming and voice over Internet Protocol (VoIP) applications, which typically have payloads of less than 100 bytes.

Packet payload compression schemes can also increase airlink efficiency. V.42bis data compression is a common compression scheme used by wireless data networks. V.42bis is based on the string compression algorithm of Ziv-Lempel [6]. V.42bis maps strings appearing in the uncompressed input data to a set of codewords. Both the compressor and decompressor dynamically construct a dictionary mapping codewords to strings. By sending shorter-length codewords over the airlink instead of the longer length strings the codewords represent, V.42bis can achieve favorable compression ratios and make more efficient use of the airlink. V.42bis, however, may be used only over links providing lossless, in-sequence packet delivery.

4.2.4. Mobility Management. Wireless data networks employ several mobility management techniques to give mobile terminals access to the Internet. Some wireless data network technologies leverage the use of mobility management architecture already deployed for the service of circuit-switched traffic (e.g., GPRS/EGPRS, UMTS). Other wireless data technologies (e.g., CDMA2000) have opted instead for mobility management based on mobile IP.

4.2.4.1. Mobility Management in GPRS/EGPRS and UMTS. Figure 5 shows the high-level mobility management architecture employed by GPRS/EGPRS and UMTS networks. In these technologies, each user is assigned a "home" service provider, the provider who bills for their service. Mobiles are also permitted to use networks other than those owned by their home service provider, a process known as "roaming."

Mobility management in GPRS/EGPRS and UMTS networks uses the following network entities:

- Visitor Location Register (VLR). This is a database containing information on the roaming mobiles currently active in a service provider's network. The VLR contains information on the roaming current mobiles' locations, the identities of the mobiles' home service provider, information needed to authenticate users, and the capabilities of roaming mobiles and other information. This database also contains information on roaming circuit-switched customers currently using the service provider's network.
- *Home Location Register (HLR).* Similar to the VLR, this database contains information on all mobiles that receive service from the service provider. This database also contains information on the service provider's circuit-switched customers.



Figure 5. Mobility management architecture employed by GPRS/EGPRS and UMTS networks.

- *GGSN*. The GGSN is a gateway node used to mask the mobility of mobile hosts from Internet applications. All data traffic sent between a mobile host and fixed host are routed via the GGSN. The GGSN collects usage information that can be later used to bill users for wireless data service. A GGSN may access the VLR/HLR to obtain information on mobile capabilities and special features that a user has subscribed to. The GGSN tunnels network layer traffic to the SGSN that a mobile host is currently attached to.
- SGSN. The SGSN tracks a mobile as it moves from cell to cell, delivering the network-layer packets it receives from the GGSN. The SGSN may compress the header and payload of the packets it transmits. The SGSN tunnels packets it receives from the mobile host to the GGSN. The SGSN also communicates with the VLR/HLR to retrieve mobile capabilities and update location information and mobile state. The SGSN also collects mobile-specific usage data that can be later used for billing.

Mobility management "procedures" are used to activate or deactivate data service and track mobiles as they move from cell to cell. During each procedure, signaling messages are exchanged between the mobile host and network entities:

- *Attach.* Before being able to transfer data over a wireless data network, a user must "attach." During the attach procedure, signaling messages are exchanged between the mobile host and the SGSN to identify and authenticate the mobile. Credentials supplied by the mobile may be compared with credentials the SGSN retrieves from the mobile's HLR. The SGSN updates the mobile's HLR with its new location. If the state information in the HLR shows that the mobile was attached with another SGSN, the new SGSN informs the old SGSN that the mobile will no longer need service from the old SGSN. The SGSN assigns the mobile a temporary link layer address which will be used to identify the mobile for as long as it remains attached.
- *Detach*. The detach procedure may be initiated by a mobile (called an "explicit detach"), or initiated in response to a lack of routing area update messages (called an "implicit detach"). The detach procedure informs the wireless data network that a mobile is no longer available to send and receive data over the wireless data network. During a detach, the VLR is updated to indicate that the mobile is now idle.
- *Packet Data Protocol (PDP) Context Activation.* PDP context activation makes a mobile "visible" to the public Internet. Signaling messages sent between the mobile and the SGSN identify the type of service that the mobile is requesting. The SGSN may perform optional procedures to determine whether the mobile is allowed access to the type of service it is requesting. If the SGSN determines the context activation should be allowed, it informs a GGSN to create a PDP context for the mobile. Once a PDP context has been created,

IP traffic can flow between a mobile and the public Internet.

- Packet Data Protocol (PDP) Context Deactivation. PDP context deactivation is used to signal the end of a data session. Once a mobile's PDP context has been successfully deactivated, the SGSN and GGSN that were serving the mobile are free to assign network resources (memory, link bandwidth, etc.) to other mobiles. At the end of PDP context deactivation, a mobile is no longer reachable from the public internet.
- Routing Area Update. To help manage the amount of signaling traffic needed to track a mobile as it moves from cell to cell, contiguous clusters of cells are grouped together to form a "routing area." Mobiles are required to inform the SGSN any time they begin to receive service in a cell with a new routing area identifier, a process known as a routing area update. Routing areas are also used to control the amount of paging traffic that must be generated to deliver traffic to mobiles in standby mode. Paging messages are only sent in those cells belonging to the routing area the mobile was last known to be in. Routing area updates are sent periodically by mobiles in standby mode. If a mobile does not send periodic updates, the network will detach the mobile, freeing up resources for other mobiles.
- *Paging*. Constant reception and decoding of airlink channels drains a mobile's battery. To increase battery life, during times of inactivity, mobiles enter a "standby mode." While in standby mode, mobiles periodically decode paging channels to determine whether the network wished to send traffic to them. Periodic decoding of paging channels can substantially increase batter, life, since data transfer tends to be sporadic.

Once a mobile host has attached and activated a PDP context, it is able to exchange IP packets with the public Internet. A "ping" [Internet Control Message Protocol (ICMP) echo] message sent from a mobile host to a fixed host is carried over the radio access network to the SGSN currently serving the mobile. The source address of the IP packet carrying the ping message is the IP address assigned to the mobile during PDP context activation. The SGSN tunnels the ping message to the GGSN currently serving the mobile's PDP context. The GGSN passes the ping message to the public Internet, where traditional IP routing is used to deliver the message to the fixed host. The fixed host echoes the ping message back to the mobile host. The destination IP address used by the fixed host is the IP address assigned to the mobile host during context activation. IP packets using the IP address of the mobile host are routed using traditional IP routing to the GGSN serving the mobile host's PDP context. The GGSN tunnels the IP packet to the SGSN serving the mobile host. The SGSN then forwards the packet to the cell currently being used by the mobile host.

4.2.4.2. Mobile IP. Mobile IP is a protocol defined by the Internet Engineering Task force to deliver



Figure 6. Mobile IP mobility management architecture.

network layer packets to mobile hosts [7]. Here we highlight a few key differences between the mobile IP mobility management approach and the approach used in GPRS/EGPRS and UMTS networks.

Mobile IP high-level mobility management architecture is shown in Fig. 6:

- Foreign Agent. The foreign agent plays a role analogous to the SGSN in GPRS/EGPRS/UMTS networks. The foreign agent is responsible for sending and receiving network layer packets to the mobile host.
- *Home Agent*. The home agent plays a role analogous to the GGSN in GPRS/EGPRS/UMTS. Data packets sent to a mobile host are first routed to the home agent. The home agent then tunnels the packets to the foreign agent currently serving the mobile host. The foreign agent "detunnels" packets forwarded by the home agent and delivers them to the mobile host.

Mobile IP employs a technique known as "triangular routing" to transfer packets to and from mobile hosts. Network-layer packets sent from a corresponding host to a mobile host, are always routed via the home agent. However, packets sent by a mobile host to the corresponding hosts bypass the home agent, and are routed directly to the corresponding host. This is in contrast to the route traveled by packets sent by mobile hosts in GPRS/EGPRS/UMTS, where packets are always routed via the GGSN, regardless of the direction they are being sent—so-called bidirectional tunnels.

4.2.5. TCP over Wireless. TCP is the predominant wireline transport layer protocol used by internet applications [8]. TCP was designed to guarantee end-toend reliable data delivery and end-to-end flow control over wireline networks. One key assumption underlies the design of TCP's flow control and error recovery mechanisms—packet loss and delay are caused primarily by network congestion. In wireless data networks, however, packet loss and delay are caused primarily by airlink transmission errors. TCP flow control and error recovery mechanisms can cause severe end-to-end performance degradation when used over wireless links. Two problems are encountered when TCP is used on wireless data networks. TCP has its own ARQ mechanism that often conflicts with the ARQ mechanism employed by a wireless data network's RLC layer. TCP segments are typically divided into several RLC PDUs, which are then scheduled for transmission. When a wireless link is bad, many retransmissions occur in the RLC layer, increasing the delay experienced by individual TCP segments. If a TCP segment is not received within certain period of time, the TCP transmitter assumes that the segment is lost and retransmits it. However, the retransmission of the TCP segment is not necessary because the first TCP segment was not lost, but delayed as a result of airlink errors. These spurious TCP retransmissions waste airlink bandwidth and reduce TCP throughput.

Another problem with the use of TCP over wireless data networks is TCP's end-to-end flow control mechanism. When TCP sees that a segment is delayed too much or lost, TCP "thinks" that there is congestion on the path between the transmitter and the receiver. Therefore, TCP reduces the transmission rate. When TCP receives a segment, it may think that the congestion has disappeared and then increases the transmission rate. However, the reason for packet delay fluctuation and loss is due mainly to the fluctuation of wireless link qualities and their induced network reactions. The way in which TCP reacts to packet delay and loss-quickly reducing transmission rates in response to excessive delay, and slowly increasing transmission rates when the delay decreases — is not well suited to the delays observed over wireless links. As a result, TCP transmission rate reductions may be too aggressive, and the wireless link may end up never fully utilized.

To solve the performance problems caused by using TCP over wireless data networks, the packet delay seen at the RLC layer should be made as smooth as possible. In this way, TCP will have an accurate estimate on the round-trip delay and then will not unnecessarily retransmit a packet or reduce the transmission rate. The delay jitter seen at the RLC layer may come from various sources, such as RLC retransmission and resequencing delay, connection teardown and resetup in the middle of a TCP session, and wireless scheduling. Although all of those causes may not be possible to eliminate, at least the designer should be aware of this and trade off the TCP problems against other considerations.

5. CONCLUSIONS

As a result of the interplay of each layer of the wireless data protocol stack, the need for spectrally efficient data transmission, the limitations placed by mobile terminal battery life, the design and engineering of wireless data networks offers many unique challenges. The wireless data networking technologies discussed in this article provide a variety of different approaches used to solve these problems.

BIOGRAPHIES

Wei Luo received a B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1995 and M.S. and Ph.D. degrees in electrical and computer engineering from University of Maryland, College Park, Maryland, in 1997 and 1999, respectively. Dr. Luo joined Bell Labs, Lucent Technologies in 1999 working on the performance, protocol and algorithm designs in wireless communication systems.

Kenneth C. Budka received his B.S. (summa cum laude) degree in electrical engineering from Union College in Schenectady, New York, in 1987 and M.S. and Ph.D. degrees in engineering science from Harvard University in Cambridge, Massachusetts, in 1988 and 1991, respectively. During his undergraduate and graduate studies, he was an exchange scholar at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland, and Columbia University in New York, New York. Dr. Budka joined AT&T Bell Labs in 1991 working on control, resource allocation, scheduling, and performance issues arising in second and third generation wireless voice, and data communications technologies. He was named a Bell Labs Distinguished Member of Technical Staff in 1999, and currently works as a technical manager in Lucent Technologies Bell Labs High Performance Communication Systems Laboratory. A senior member of the Institute of Electrical and Electronics Engineers, Dr. Budka holds two patents in the area of wireless voice and data networks, with more than 10 others pending.

Krishna Balachandran received his B.E. (Hons) degree in electrical and electronics engineering from Birla Institute of Technology and Science, Pilani, India, M.S. in computer and systems engineering, M.S. in mathematics and Ph.D. in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, New York. He joined Bell Labs, Lucent Technologies in 1996 as a member of technical staff and is currently an acting technical manager in the Networking Technologies and Performance Department at Bell Labs. His research interests include the design, analysis and simulation of advanced physical layer, media access control and radio link control protocols, link adaptation/hybrid ARQ techniques, resource allocation, and scheduling techniques for wireless systems. Dr. Balachandran has published over 30 journal and conference papers in related areas. He also holds three patents (several pending) and has contributed significantly to standardization activities in the TIA, ETSI, and 3GPP.

BIBLIOGRAPHY

- 1. IEEE std 802.11b-1999, Supplements to ANSI/IEEE std 802.11, 1999 ed.
- 2. 3GPP TS 25.401, UTRAN Overall Description, v.3.1.0, Jan. 2000.
- 3. 3GPP TS 23.107, QoS Concept and Architecture, v.3.2.0, March 2000.
- S. Nanda, K. Balachandran, and S. Kumar, Adaptation techniques in wireless packet data services, *IEEE Commun. Mag.* 38(1): (Jan. 2000).
- 5. V. Jacobsen, Compressing TCP/IP Headers for Low-Speed Serial Links, IETF RFC 1144, 1990.

- J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans. Inform. Theory* 23(3): 337–343 (May 1977).
- 7. C. Perkins, IP Mobility Support, IETF RFC 2002, Oct. 1996.
- 8. W. R. Stevens, TCP/IP Illustrated, Vol. 1, *The Protocol*, Addison-Wesley, Reading, MA, 1994.

WIRELESS SENSOR NETWORKS

SEAPAHN MEGERIAN MIODRAG POTKONJAK University of California at Los Angeles West Hills, California

1. INTRODUCTION

Since the early 1990s, the sustained high pace of technological advances paved the way for the exponential growth of the Internet. We can trace the development of two implementation technologies as prime enablers of this growth. The first was the dramatic reduction in the cost of disks, namely, massive long-term storage. The second was the huge reduction in the cost of optical communication and its simultaneous capacity increase. More specifically, since 1991, the capacity of a \$100 disk increased by a factor of 1200, while during the same period, the bandwidth of optical cable doubled every 9 months. The Internet, as we know it today, is an exceptional educational, research, entertainment, and economic resource, which enables information to be available at the touch of a mouse. There is a wide consensus that the Internet will continue to grow rapidly in both quantitative and qualitative terms. At the same time, it appears that we are on the brink of the next technological revolution that may have even more profound impact on our lives. This revolution, that will enable any time, anywhere, communication and connection between the physical and computational worlds, is due to the advancement of wireless communication technology and sensors. While in the early 1990s wireless technology was mainly stagnant, since 1996, it has experienced an exponential growth. Wireless bandwidth in industrial offerings has increased by a factor of 28 from 1997 to 2002. On the other hand, recent progress in fabrication of micro-electromechanical systems-based (MEMS) sensors has opened new vistas in terms of cost, reliability, accuracy, and low energy requirements. While most of the MEMS-based sensors are still in the research phase, a boom in government funding in this area has resulted in amazing progress. For this field, the total funding was \$2 million in 1991 and \$35 million in 1995, while in 2001 it was estimated to have been \$300 million worldwide. With such advancement, there is currently a need for methodologies and technologies that will enable efficient and effective use of wireless embedded sensor network applications. The motivational factors pushing for these applications include the mobility of computational devices, such as cellular phones and personal digital assistants (PDAs), and the ability to embed these devices into the physical world.

Almost all of the modern science and engineering has been built using compound experiment-theory iteration steps. Typically, the experiments have been the expensive and slow components of the iterations. Thus, the existences of flexible yet economic experimentation platforms often result in great conceptual and theoretical breakthroughs. For example, advanced optical and infrared telescopes enabled spectacular progress in the understanding of large scale cosmology theory. Particle accelerators and colliders enabled great progress in the understanding the ultra small world of elementary particles. Furthermore, the progresses in computer science, information theory, and nonparametric statistics have been greatly facilitated by the ability to compile and execute programs quickly on general-purpose computers. Sensor networks will enable the same type of progress in better understanding many other sciences, not just by information processing, but also through new connections between the sciences and the physical, chemical, and biological worlds.

Sensor networks consist of a set of sensor nodes, each equipped with one or more sensors, communication subsystems, storage and processing resources, and in some cases actuators. The sensors in a node observe phenomena such as thermal, optic, acoustic, seismic, and acceleration events, while the processing and other components analyze the raw data and formulate answers to specific user requests. The recent advances in technology mentioned above, have paved the way for the design and implementation of new generations of sensor network nodes, packaged in very small and inexpensive form factors with sophisticated computation and wireless communication abilities. Once deployed, sensor nodes begin to observe the environment, communicate with their neighbors (i.e., nodes within communication range), collaboratively process raw sensory inputs, and perform a wide variety of tasks specified by the applications at hand. The key factor that makes wireless sensor networks so unique and promising in terms of both research and economic potentials is their ability to be deployed in very large scales without the complex preplanning, architectural engineering, and physical barriers that wired systems have faced in the past. The term "ad hoc" generally signifies such a deployment scenario where no structure, hierarchy, or network topology is defined a priori. In addition to being ad hoc, the wireless nature of the communication subsystems that rely on radio frequency (RF), infrared (IR), or other technologies, enable usage and deployment scenarios that were never before possible.

To illustrate the key concepts and a possible application of wireless ad hoc sensor networks (WASNs), consider the environmental monitoring requirements of large office buildings. Such buildings typically contain hundreds of environmental sensors (such as thermostats) that are wired to central air conditioning and ventilation systems. The significant wiring costs limit the complexity of current environmental controls and their reconfigurability. Furthermore, in highly dynamic corporate environments, cubicle offices may continuously be added, removed, and restructured, which makes environmental control rewiring an intractable task. However, replacing the hard-wired monitoring units with inexpensive ad hoc wireless sensor nodes will easily improve the quality and energy efficiency of the environmental system while allowing unlimited reconfiguration and customization in the future. In addition to the classic temperature sensing, senor nodes with multiple modalities (i.e., equipped with several different types of sensors) can significantly enhance the abilities of such a system. For example, motion or light sensors can detect the presence of people and even adjust the environmental controls using actuators, according to prespecified user preferences. In many instances, the savings in the initial wiring costs alone may justify the use of such wireless sensor nodes.

Although the environmental monitoring example above is an application of WASNs to a task that has existed for a long time, many new applications have also started to emerge as direct consequences of WASN developments. Such applications range from early forest fire detection and sophisticated earthquake monitoring in dense urban areas, to highly specialized medical diagnostic tasks where tiny sensors may even be ingested or administered into the human body. As mentioned above, personal spaces such as offices and living rooms can be customized to each individual by sensors that detect the presence of a nearby person and command the appropriate actuators to execute actions according to that person's preferences. In essence, WASNs provide the final missing link connecting our physical world to the computational world and the Internet. Although many of these sensor technologies are not new, technological barriers and physical laws governing the energy requirements of performing wireless communications have limited their feasibility in the past. A few highlights and benefits of the newer, more capable sensor nodes are their abilities to

- Form very large-scale networks (thousands or more nodes).
- Implement sophisticated networking protocols as well as distributed and localized analysis algorithms.
- Reduce the amount of communication required to perform tasks by distributed and/or localized computations.
- Implement complex power-saving modes of operation depending on the environment, current tasks, and the state of the network.

In the following sections, we describe the generic components that form a wireless sensor network and highlight the key issues and characteristics that differentiate sensor networks from traditional peer-to-peer and ad hoc wireless communication networks. Section 2 lists the architectural and hardware related components, while in Section 3 the focus is on higher-level services and software issues. Section 4 provides a brief overview of the state of the art and the challenges ahead.

2. ARCHITECTURE AND HARDWARE

Similar to classical computer architectures, the main components of the physical architecture of WSN nodes can be classified into four major groups: (1) processing, (2) storage, (3) communication, and (4) sensing and actuation [input/output (I/O)]. The following is a brief summary of the main issues involved and some related topics for each of these components.

2.1. Processing

Two key constraints for processing components are energy and cost. Essentially all current WSN processors are those used for mass markets. This is due in large part to the advantages of the economies of scale and the availability of comprehensive and mature software development environments for such processors. Since the processing in a node has to address a variety of different tasks, many nodes have several types of processors: microprocessors and/or microcontrollers, low-power digital signal processors (DSPs), communication processors, and application-specific integrated circuits (ASICs) for certain special tasks. The standard complementary metal oxide semiconductor (CMOS) process will be the technology of choice for sensor node processors at least until 2012.

2.2. Storage

Currently, sensor nodes have relatively small storage components. They most often consist of standard dynamic random access memory (DRAM) and relatively large quantities of nonvolatile (flash) memory. Since the communication is a dominant component of the overall energy consumption is wireless sensor networks, we expect that the amount of local storage at a node will continue to increase. This expectation is further enforced by the fact that since 1992, the cost of memory was declining much faster than the cost of processors. We also expect that new technologies, in particular magnetoresistive randomaccess memory (MRAM), will soon be widely used for this type of storage.

2.3. Communication

The communication paradigms often associated with the current generations of wireless sensor networks are multihop communication. Several current results indicate that multihop communications scale very well and can significantly reduce the energy consumption in large sensor networks [1]. A number of new projects are currently targeting low power communication. This is an area where it is most difficult to predict how technology will impact future architectures, since commercial wireless communication is a relatively new field. It is very important to note here that in typical low-power radios used in WASN communication, listening often requires as much energy as transmitting. This is in sharp contrast to the assumptions made in most previous work in ad hoc multihop networking, where sending a message was believed to have been the major consumer of energy. These new constraints indicate that the study of complex power saving modes of operation, such as having multiple different sleep states, will be crucial in this field.

2.4. Sensors and Actuators

One can envision the sensors as the eyes of the sensor network, and the actuators, as its muscles. Although MEMS technology has been making steady progress since the early 1960s, it is obvious that it is still in its early phases where development is mainly sustained by research funding and not yet commercial. However, significant results have already been obtained. A good starting point for learning more about sensor systems is the article by Mason et al. [2].

3. SYSTEM SOFTWARE AND APPLICATIONS

As described above, the recent advent of WASNs has required completely new approaches for building system software and optimization algorithms, as well as the adaptation of existing techniques. It is interesting and important to analyze why the already existing distributed algorithm techniques were not directly applicable to WASNs. There are at least five major reasons: (1) WASNs are intrinsically related to the physical and geometric world and therefore have very special properties-the uses of local and geographic information, for example, play key roles in designing efficient, robust, and scalable sensor networks; (2) relative communication costs are much higher than they were assumed to be in all previous distributed computing research - since WASN nodes are severely energy constrained, the cost of communication becomes an extremely important factor in the design of WASN software; (3) accuracy of physical measurements is intrinsically limited and therefore there is little advantage on insisting on completely accurate results; (4) energy consumption is a critical system constraint; and (5) data acquisition is naturally distributed and error-prone, implying a strong need for new sensing, computation, and communication models.

The relative communication delay in sensor networks is significantly larger than in traditional computational systems. It is interesting to note that in modern deep submicrometer (DSM) chip designs, delay on a single system-on-chip will be up to 20 clock cycles. However, even the fastest communication protocols in WASNs will have delays in millions of cycles due to technological and physical limitations as well as system software overhead. Furthermore, communication generally dominates both sensing and computation in terms of energy (currently, image and video sensors are exceptions). Again, it is interesting to draw parallels with DSM designs: In DSM, communication will also dominate power consumption, maybe eventually by as high as a 10:1 ratio with respect to computations. In WASNs, technology trends are much more difficult to predict, yet at least in current and pending technologies, this ratio is much higher, often estimated at 1000:1.

Interestingly, several new hardware and architectural characteristics have also come into play that strongly influence WASN communication costs. For example, we have already mentioned that in many of the current low-power radios used in WASN nodes, the power requirements for listening or receiving messages is about the same as when transmitting. This is in sharp contrast to the assumptions made in numerous wireless communication research efforts in the past, where transmitting a message was almost always assumed to have required much more power than listening or receiving a message. Consequently, in order to be truly effective, WASN system software must try to maximize the duration of the times when the communication subsystems can be turned off or placed in sleep modes, thus saving precious reserve energy resources.

In addition to placing nodes in sleep modes to conserve energy, one can expect that fault tolerance and autonomous operation will be essentially mandatory for large scale WASNs, due to wide-scale deployment and the relatively high cost of servicing nodes. During the useful lifetime of a typical WASN, it is not unreasonable to expect that at least some nodes will exhaust their energy supply. Even if latency (real-time constraints), energy consumption, and fault tolerance were not an issue, security and privacy issues would very often mandate that only a subset of nodes participate in a specific task. In addition, sensor nodes are often deployed outside strictly controlled environments, communicate using wireless (insecure) media, and hence are highly susceptible to security attacks. This further indicates that expecting all nodes to always be able to sense, communicate, and compute is not realistic. Moreover, as WASNs evolve into an Internet-like scale and organization and span the whole earth and beyond, the only realistic possibility for all tasks will be execution in highly localized scenarios. In localized computation models, only a subset of nodes, which are almost always within geographic proximity, collaborate and participate in formulating results to specific application tasks.

The challenges outlined above can be classified into three major categories: (1) strict constraints; (2) new modes of operation; and (3) interface between physical world, computation, and information theory. The "strict constraint" challenges include problems related to the need for low cost, long life, and reliable infrastructures. Low-power operation, wireless bandwidth efficiency, reliability, fault tolerance, high availability, error recovery, distributed synchronization, and realtime operation in unpredictable environments are all important factors that influence the design decisions at this step. In this direction, the current key problem is learning how to scale the already available techniques to the next levels of strictness of constraints.

There are two main research direction related to the "new modes of operation" of WSNs, due to their distributed and multihop natures: localized algorithms and autonomous continuous operation. *Localized algorithms* are algorithms implemented on sensor networks in such a way that only a limited number of nodes communicate, therefore reducing overall energy consumption and bandwidth requirements. Consequently, localized algorithms often operate with incomplete information, noisy data, and almost always under very strict communication and energy constraints.

One way of modeling localized algorithms in WASNs is as follows: One or more nodes initiate a request for a computation (a query). The result of the query is to be sent to one or more sink nodes. Each node can obtain its required information either through its sensors or by communication with neighboring nodes. The goal is to maximize an objective function for the optimization task at hand, in such a way that all constraints are satisfied and the communication cost is minimal. The first and most important difference between the localized algorithm and other traditional methods is the amount of information available to the processing units. In conventional scenarios, the processors have all the information that is needed for their computation tasks. However, in localized approaches, the required information is not complete and thus the communication between components should be interleaved with the computations in different parts so that they compensate for the insufficient information. The other interesting aspect of localized procedures is that although there are many processing units in a pervasive computing environment such as WASNs, for most of the applications, only a few processors are sufficient to carry out the required calculations. This is in contrast to the classical distributed computing paradigms where all processors involved in a computation are actively computing all the time. For centralized algorithms, of course, one processing unit handles all the computations and control. In addition to the localized nature of the optimization algorithms in WASNs, autonomous closed-loop modes of operation are a must for effective use of such networks. Essentially, the applications must execute with minimal or no intervention of a human operator.

Traditional wired and wireless computer communication network designers have typically followed (although often loosely) the International Organization for Standardization (ISO¹) Open System Interconnection (OSI) Reference Model as the basis for their protocol stack design. The OSI Reference Model specifies seven protocol layers: physical, data link, network, transport, session, presentation, and application [3]. The following subsections briefly describe two main WASN protocol stack layer functions, namely, medium access control (MAC) and routing, which are equivalent to what the OSI model classifies as data-link- and network-layer functions, respectively. The subsequent sections then describe sensor network specific tasks and problems such as location discovery and coverage.

3.1. Medium Access Control (MAC)

Wireless communication media are almost always broadcast in nature and thus are shared among the participants. For example, RF transmissions of one node can be "heard" by any other node that is within communication range. If two nodes that are close together transmit at the same time, their transmissions will most likely "collide" and interfere with each other. *Medium access*

 $^{^1\,\}mathrm{ISO}$ is not an acronym. ISO is an international standardization organization with members from more than 75 countries.

control refers to the process by which nodes determine when and how to utilize a shared communication medium. In WASNs where network communications are multihop (often require intermediate nodes to forward packets), the MAC layer is also where specific self-organization and autonomous configuration abilities can be introduced into the network.

Traditional MAC designs have followed two distinct philosophies: dedicated and contention-based. In the dedicated scenarios, each node receives the shared resource according to a prespecified scheme. Time-division multiple access (TDMA) is one such scheme where each node may only transmit within a small, periodic, time slot. Such MAC strategies are typically not well suited for ad hoc networks that have no predefined organization and can be very dynamic in nature; that is, nodes can join, move, or leave the network at any time. In contention-based schemes, nodes attempt to "grab" the medium and transmit when needed. Often, nodes have abilities to sense that a channel is in use and thus determine that they must wait. References 4 and 5 provide an overview of existing techniques and propose new MAC layer schemes that are designed specifically for WASNs.

3.2. Ad Hoc Routing

Routing refers to the process of finding ways to deliver a message from a source to its destination. In ad hoc, multihop networking scenarios, routing is an especially difficult problem since the nodes must discover the destination and the routes to the destinations subject to extreme energy consumption limitations. Existing works from ad hoc wireless networking domains provide a solid foundation for WASN routing problems. However, WASNs have unique features that make traditional routing philosophies less relevant. In traditional wired and wireless data communication networks, connections are peer-to-peer. This means that the user at a specific source node must send data (usually in forms of packets) to another user at a specific destination. Consequently, the endpoints of communication typically have unique names and specific communications are identified by the source and destination names (addresses). In WASNs, however, such peer-to-peer communications are less meaningful. Typically, nodes that sense events, analyze the data, collaborate with neighbors, and communicate processed information to one or more sink nodes. In addition, the information may be processed further along the path to the destination which makes the definition of "routing" very vague in WASNs compared to traditional data communication networks.

"Flooding" is a well-known basic scheme that can be used for routing in any network. During flooding, each intermediate node that receives a packet simply forwards it to all its neighbors until it reaches the destination. In a connected network, the packet will most likely reach the destination, although packet losses due to interference and other transmission errors are always possible. Although for broadcast messages flooding is very effective, the overhead for point-to-point communication is extremely high. Other more sophisticated approaches have been proposed such as dynamic source routing (DSR) and ad hoc on-demand distance vector (AODV) routing, which try to discover routes and maintain information about the network topology to eliminate flooding overheads. Reference 6 provides an overview and detailed analysis of several ad hoc network routing protocols. However, as stated above, routing schemes from ad hoc networking do not necessarily work well in sensor networks.

Several schemes have been proposed for routing in WASNs that leverage on sensor network specific characteristics such as geographic information and application requirements. Because of the immaturity of the field, none have established themselves as definitive solutions to WASN routing. Directed diffusion is one example of a generic scheme for managing the data communication requirements (and thus routing) in WASNs. The basic scheme in directed diffusion proposes the naming of data as opposed to naming sources and destinations of data. Data are "named" using attribute-value pairs. Data are requested by name as "interests" in the network. The request (dissemination) sets up "gradients" so that the named data (or events) can be "drawn." In traditional IP-style communication, nodes are identified as "endpoints" and the communication is layered as an end-to-end service. In directed diffusion, in contrast, named data flow toward the originators of their corresponding interests along multiple paths with the network "reinforcing" one or multiple such paths [7]. However, as stated above, WASN nodes may process the data at intermediate steps and the specific routing solution may be tightly coupled with application tasks (as opposed to layered).

3.3. Location Discovery

Geographic information is an integral attribute of any physical measurement. Thus, the knowledge of node locations is fundamental in proper operations of sensor networks, especially for WASNs. The ad hoc nature of WASN deployment necessitates that each node determine its location through a location discovery process. The Global Positioning System (GPS) is one method that was designed and is controlled by the United States Department of Defense for this purpose. The GPS system consists of at least 24 satellites in orbit around the earth, with at least four satellites viewable from any point, at a given time, on earth. They each broadcast time-stamped messages at periodic intervals. Any device that can hear the messages from four or more satellites can estimate its distance from each satellite and thus perform trilateration to compute its position.

Although GPS is an elegant solution to the location discovery process, it has several limitations that hinder its use in WASN applications: (1) GPS is costly in terms of both hardware and power requirements and (2) it requires line-of-sight communication between the receiver and the satellites and thus does not work well when obstructions such as buildings, trees, and mountains block the direct "view" to the satellites. Thus, other techniques have been proposed to dynamically compute the locations of the nodes in WASNs. In several location discovery schemes, the received signal strength indicator (RSSI) of RF communication is used as a measure of distance between nodes. In other schemes, the time difference in arrival of RF and acoustic (ultrasound) signals are used to approximate node distances. Once nodes in a WASN have the ability to estimate distances between each other (ranging), they can then compute their locations using the simple trilateration method. In order for a trilateration to be successful, a node must have at least three neighbors who already know their locations. This requires that at least a subset of nodes determine their locations through other means such as by using GPS, manual programming, or deterministic deployment (placing nodes at specified coordinates). References 8 and 9 provide detailed discussions on location discovery techniques and algorithms.

3.4. Coverage

Several different coverage formulations arise naturally in many domains. The "art gallery problem," for example, deals with determining the number of observers necessary to cover an art gallery room such that every point is seen by at least one observer. This problem has several applications such as for optimal antenna placement problems in wireless communication. The art gallery problem was solved optimally in two dimensions (2D) [10] and was shown to be computationally intractable in the 3D case. Coverage in the context of sensor networks can have very new semantics. The main question at the core of coverage is trying to answer how well the sensors observe a physical space. References 11 and 12present several formulations of sensor coverage in sensor networks. The formulations include calculations based on best- and worst-case coverages for agents moving in a sensor field and exposure-based methods. In the bestand worst-case formulations, the distance of the agent to the closest sensors are of importance, while in exposurebased methods the detection probability (observability) in the sensor field is represented by a path-dependent integral of multiple sensor intensities. In both of these schemes, the types of actions that the agent performs impact the coverage metric. For example, the sensor field may have a different coverage level if an agent is traveling west to east as opposed to north to south, or along any other arbitrary paths. The actual physical characteristics and abilities of WASN nodes will play crucial roles in building practical, accurate, and useful coverage models and analysis algorithms.

4. FUTURE DIRECTIONS

We conclude by summarizing some important future challenges in wireless sensor networks:

QoS. For quality of service (QoS), one can define both syntactic and semantic interpretations. On the syntactic level, one can consider dimensions such as coverage, exposure, latency, measurement and communication errors, and event detection confidence. On the semantic level, one can define utility and cost functions to enable the analysis of how particular data can help in the construction of more accurate models of the physical world or more efficient algorithms.

- Scaling. Scaling has been the key metric in analyzing both graph-theoretic and physical phenomena. The goal will be to develop new methods that are based on statistical techniques instead of traditional probabilistic ones. Existing techniques such as state transitions and percolation will be key factors in analyzing and building very large systems and optimizing their performance.
- Profilers, Recommenders, and Search Engines. Profilers, recommenders, and search engines rapidly emerged as mandatory systems enabling efficient use of the World Wide Web (WWW) and the Internet. There are clear needs to develop such systems for sensor networks. New dimensions and challenges include ways to include information and knowledge, not just about physical location and physical time, but also about the physical, chemical, and biological worlds. There are needs for profilers of events, objects, areas, sensors, and users, among other things.
- Foundations and Theory. There is a need to develop new theoretical foundations, new models, new algorithmic complexity theory and practice, new programming models, and languages for embedded sensor networks. For example, new models of sensor networks will encompass the already existing Markov models, interacting particle models (e.g., the Ising model), bifurcation-based models, fractals, oscillations, and space pattern models. In addition, there will be a need to create new models unique to wireless sensor networks. As another example, the VLSI (very large scale integration) theory field was built based on two lasting premises: (1) that integrated circuits are planar and (2) that features are of small size, yet limited in quantity. There is a need to explore such lasting features in sensor networks. Examples of such rule-based modeling are "energy spent on communication is dominant and distance-dependent," "all measurements have intrinsic errors," and "storage space on nodes is very limited."

Other potential research directions include validation and debugging, data compression and aggregation, realtime constraints, distributed scheduling and assignment, pricing, and privacy of actions.

BIOGRAPHIES

Seapahn Megerian (Meguerdichian) is currently a Ph.D. student in the Computer Science Department at the University of California, Los Angeles. He received his Computer Science and Engineering B.S. degree in 1998 and Computer Science M.S. degree in 1999 from UCLA. His primary focus is in the design and development of efficient algorithms for deployment,

2996 WIRELESS SENSOR NETWORKS

performance, and coverage analysis; decision support; operation optimization; security; and privacy in wireless ad hoc sensor networks. In addition, his research includes high-performance communication systems, system-onchip network design, application-specific compilers, and computational security. He was the recipient of the 7th Annual International Conference on Mobile Computing and Networking (MobiCom 2001) Best Student Paper Award for the paper titled "Exposure in Wireless Ad Hoc Sensor Networks."

Miodrag Potkonjak is a Professor in the Computer Science Department at the University of California, Los Angeles. He received his Ph.D. degree in Electrical Engineering and Computer Science from University of California, Berkeley in 1991. In 1991, he joined C&C Research Laboratories, NEC USA in Princeton, New Jersey. Since 1995, he has been with UCLA. He has received the NSF CAREER award, OKAWA foundation award, UCLA TRW SEAS Excellence in Teaching Award, and a number of best-paper awards. His research intersets include complex distributed systems, communication system design, embedded systems, computational security, practical optimization techniques, and intellectual property protection.

BIBLIOGRAPHY

- 1. J. M. Rabaey et al., PicoRadio supports ad hoc ultra-low power wireless networking, *Computer* **33**: 42–48 (July 2000).
- A. Mason et al., A generic multielement microsystem for portable wireless applications, *Proc. IEEE* 86: 1733–1745 (Aug. 1998).

- 3. http://webopedia.internet.com/quick_ref/OSI_Layers.html.
- 4. W. Ye, J. Heidemann, and D. Estrin, An energy-efficient MAC protocol for wireless sensor networks, *IEEE Infocom* (in press).
- A. Woo and D. Culler, A transmission control scheme for media access in sensor networks, Proc. ACM/IEEE Int. Conf. Mobile Computing and Networking (MobiCOM 2001).
- J. Broch et al., A performance comparision of multi-hop wireless ad hoc network routing protocols, *Proc. ACM/IEEE Int. Conf. Mobile Computing and Networking*, Oct. 1998, pp. 85–97.
- C. Intanagonwiwat, R. Govindan, and D. Estrin, Directed diffusion: A scalable and robust communication paradigm for sensor networks. Proc. 6th Annual Int. Conf. Mobile Computing and Networking (MobiCOM 2000), 2000, pp. 56–67.
- P. Bahl and V. N. Padmanbhan, RADAR: An in-building RFbased user location and tracking system, *Proc. IEEE Infocom* 2000, April 2000, pp. 775-784.
- A. Savvides, C. C. Han, and M. B. Srivastava, Dynamic finegrained localization in ad-hoc networks of sensors, *Proc.* 7th Annual Int. Conf. Mobile Computing and Networking (MobiCOM 2001), July 2001.
- J. O'Rourke, Computational geometry column 15 (open problem from art gallery solved), Int. J. Comput. Geom. Appl. 2: 215–217 (June 1992).
- S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. Srivastava, Coverage problems in wireless ad-hoc sensor networks, *IEEE Infocom 2001* 3: 1380-1387 (April 2001).
- S. Meguerdichian, F. Koushanfar, G. Qu, and M. Potkonjak, Exposure in wireless ad-hoc sensor networks, *Proc. 7th* Annual Int. Conf. Mobile Computing and Networking (MobiCOM 2001), July 2001, pp. 139–150.

Note: Boldface numbers indicate illustrations and tables.

A erlangs, traffic engineering and, 488 A law companders, 529, 529 a posteriori probability algorithm continuous phase modulation and, 2180-81 finite geometry coding and, 805 low density parity check coding and, 1313, 1316 serially concatenated coding for CPM and, 2180-81 speech coding/synthesis and, 2367 tailbiting convolutional coding and, 2515 threshold coding and, 2580, 2584 trellis coding and, 2648, 2650 turbo coding and, 2705, 2714 a priori error, acoustic echo cancellation and, 8 absolute threshold, speech coding/synthesis and, 2364 absorbing boundary conditions, antenna modeling and, 177 absorption, 2065 free space optics and, 1851, 1855-57, 1856 lasers and, 1776 microwave and, 2560 millimeter wave propagation and, in clear air, 1270, 1270, 1436-37, 1439 optical fiber and, 1709, 1710 photodetectors and, 995, 995 absorption fading, 2065 access control (see also admission control), 1153, 1648, 1649-51 cdma2000 and, 366-367 discretionary, 1649 mandatory, 1649 role-based in, 1649 satellite onboard processing and, 482 access networks home area networks and, 2688, 2688 optical fiber systems and, 1840, 1841 powerline communications and, 1997-99, 1997 virtual private networks and, 2707-08 access points satellite communications and, 2117 wireless communications, wireless LAN and, 1285 access time CDROM and, 1735 hard disk drives and, 1321 ACES satellite communication, 196 ACH algorithm, constrained coding techniques for data storage and, 578, 579, 581 acknowledgment, 226, 345 acoustic communications, underwater (see underwater acoustic communications) acoustic communications (acomm) (see also acoustic modems for underwater communications), 15 Acoustic Communications Advanced Technology Demonstration, 24 acoustic echo cancellation, 1-15, 1 a priori error in, 8 adaptive algorithms for, 7-9 adaptive filters in, 6 affine projection algorithm for, 7-8, 8 attenuation in, 4-5 block processing solutions for, 12 center clipper in, 1, 1 decorrelation filters in, 7, 8 delay coefficients and, 10 doubletalk and, 4, 11 earliest use of, 1 echo cancellation filter in, 1, 2, 2, 3, 4, 5, 6, 8, 12 echo return loss enhancement in, 3 electronically replicated LEMS and, 3 filters in, 1, 2, 6, 7, 8 FIR filter in, 3, 3 forgetting factor in, 8, 9 full- to half-duplex connection in, 5 IIR filter in, 3 image source in, 3 impulse response of LEMS and, 2, 2, 3 input signal matrix for, 7

loss control circuit in 1 2 loudspeaker enclosure microphone system in, 1, 2, 2-3, 6 mismatch vector in, 6, 7 noise and, 4, 5 normalized least mean square algorithm for, 7, 8, 9 - 12power comparison estimation for, 10-11 power spectral density in, 6 public address systems and, 6 recursive least squares algorithm for, 8-9, 8 reflection coefficient in, 3 residual echo suppressing filter in, 1, 2, 5-7 reverberation time in, 2 shadow filters and, 10 side constraints in, 4-5 signals and, 4, 5 simulated LEMS and, 3 singletalk and, 4 speech signals and, 4, 5 stabilizing electroacoustic loop in, 5-7 standards, 4 step size control of NLMS algorithm in, 7, 9-11 stereophonic systems and, 11, 11 subband, 11-12, 11 switching for, 5 system distance in, 6 undisturbed error in, 6, 10 Wiener equation in, 6 acoustic jitter, solitons and, 1767 acoustic modems acoustic telemetry in, 23-24 Naval Undersea Warfare Center range-based modem in, 25–26, 25 underwater communications using, 15-22 acoustic modems for underwater communications, 15-22 acquisition waveforms in, 16 analog to digital converter in, 17-18 applications for, 19-22 automatic gain control (AGC) and, 17, 18-19 bandwidth in, 15, 16 coherent vs. noncoherent processing in, 16-17 cone penetrometer using, 19-20, 20 datalogging in, 18 digital signal processors in, 17, 17 digital to analog converter in, 17 Doppler shift and, 18-19 fading in, 15 hardware implementation in, 17-18 imaging and telemetry using, 20-21, 21 intersymbol interference (ISI) in, 15 modulation in, 16, 19 multipath interference in, 15 OSI reference model networks and, 15 packet data in, 15-16 pipeline bending using, 20, 20 propagation delay in, 15 RS-xxx interfaces for, 18 serial interface for, 18 signal processing in, 18-19, 19 signal to noise ratio in, 15-17 speed of sound in water and, 15 transmission loss in, 15 transmitter/receiver network for, 17 acoustic telemetry, 22-29 acoustic modems for, 23-24 applications for, 24-27, 26, 27 bandwidth and, 22 digital signal processor and, 23 direct sequence spread spectrum and, 23, 27 Doppler shift and, 23 mobile deep range, 26, 26 modulation in, 23, 24 multipath interference and, 22 Naval Undersea Warfare Center range-based modem in, 25–26, 25

range-based, applications for, 26-27, 26, 27 receivers for, 23, 23 research in, 27-28 signal to noise ratio and, 22 synchronization signals in, 23 synthetic environment tactical integration visual torpedo program and, 26, 27 transmitter for, 23, 23 turbo equalization, turbo coding and, 28 underwater applications for, 22-29 underwater range data communications and, 26-27 acoustic transducers, 29-36 attenuation and, 30, 30 ceramics in, 34, 35 condenser microphone using, 34, 34 density of media, sound propagation in, 30-31, 31 development of, 29 electro-, 29 equivalent circuits and, 29 examples of, 34-35 flexural air ultrasonic type, 35, 35 gas vs. liquid media, sound propagation in, 30-31, 31 materials for, 34 moving coil electrodynamic loudspeaker using, 34, 34 Ohm's law analogy to, 31, 31 particle displacement and particle velocity in, 31-32 propagation of sound, sound generation in, 29-30, 30 radiation patterns in, 32-33, 33 resonance in, 33-34 sonar and, 29 sound pressure and, 31 sound pressure level and, 32 speed of sound and, 30 telephone and, 29 tonpilz sonar and, 35, 35 transduction in. 34 wavelength of sound and, 30, 30 acoustoopic filters, 1729-30, 1756 acoustooptical gratings, 1755-56, 1756 acoustooptical tunable switches, 1785 acquisition waveforms, modem, 16 acquistion and tracking, IS95 cellular telephone standard and, 354 activation function, neural networks and, sigmoidal, 1676 active antennas, 47-68 advantages of, 53 amplitude modulation and, 49, 50 applications of and prospects for, 64-66 array factor in, 62–63, 62 brightness theorem and, 65 capacitance and, 49-50, 55 capacitive impedance and, 49 Cartesian coordinate systems and, 61 coplanar waveguide, 51-52, 52, 64-66 coupling and, 52 damping and, 60 dipole, 48 directivity effect in, 63 Doppler sensors and, 65 electrical fields and, 61 energy density equations for, 54 equivalent circuits and, 60-61, 61 feedback networks and, 58-59, 59 frequency modulation and, 50 Fresnel coefficient in, 56 gain in, 58 grid oscillators and, 66, 66 Hertz type, 48-68 impedance and impedance matching in, 49, 50, 50, 56 inductance in, 55 linear arrays for, 62-63, 62 loading element use of, 64-65, 64 locked beam, 63-66

active antennas (continued) lossy vs. lossless transmission lines and, 55-56, 56 magnetic flux and, 55 Maxwell's equations for, 53-54 microstrip line, 51-52, 52, 62, 62 microwave communications and, 47-49, 65 modulation and, 49 open circuit gain in, 58 open waveguide design, 51-52 oscillators and, 52 oscillators and feedback circuits and, 51, 51, 52, 58-60, 58, 59, 60, 63-66, 66 parallel plate capacitor and, 49–50, **49** patch antennas and, 52–53, **52**, **53**, 61–62, **61** phase conjugates and, 65 planar arrays and, 61 power handling in, 53 power handling in, 65-66, 65 Poynting vector in, 53-57, 61 proximity detectors using, 65 quantitative aspects of, 53-64 quasioptics in, 53 quasistatistic approximation and, 54 radar and, 51 resistive impedance and, 49 resonance and, 48 retroreflection and, 65 RLC networks and, 64 shortwave, 49 steering in, 66 tank circuits and, 62 television, 50 transfer electron devices in, 51 transistors and FETs in, 51, 57-58, 57, 58, 65 transmission lines and, 50-55, 50 transmitters and, 50-51, 51 voltage reflection coefficient and, 56 voltage waves and, 56 wavelength and, 49, 49 active attacks in. 1646-47 active integrated antenna, 1429-31, 1430, 1431 active layer, lasers and, 1777 active phase aray antenna, 1391, 1391 active queue management, 1661 ad hoc communications systems, networks, 308, 309, 1285, 2883–99, 2994 ad hoc on demand distance vector, 2211, 2887-88, 2891, 2894 ADAPT protocol, media access control and, 0, 1348 adaptation algorithm, adaptive equalizers and, 82 adaptive algorithms, 7-9, 68 adaptive antenna arrays, 68-79, 163, 180, 184, 186-187, 192, **192** absence of mutual coupling in, 73-74, 74 adaptive algorithms used in, 68 beamforming, 187 beamwidth in, 187 cochannel interference and, 454-455 coding division multiple access and, 187 conjugate gradient method in, 72-73 coupling in, 68-70, 73-74, 74, 74-77, 75, 76, 77 covariance and covariance matrix in, 68 degrees of freedom calculations in, 73 direct data domain least squares method in, 71-73 eigenvalues for, 72 error criterion in, 72-73 gain and, 68 interference and, 68, 69-71 jamming and, 74, 74 least mean squares algorithm in, 69, 71-73 Maxwell's equations for, 70 mean square error in, 187 method of moments in, 68 noise and thermal noise in, 71-72, 74, 74 numerical examples for, 73-77 presence of mutual coupling in, 74-77, 75, 76, 77 satellite onboard processing and, 479-480 scatter in, 68-69 semicircular array in, 75-77, 75, 76 signal processing and, 68 signal to noise ratio (SNR) in, 187

steering vectors in, 69, 70 transformation matrix in, 70-71, 70 uniform linear virtual array in, 69, 70-71, 75-77 adaptive delta modulation, 2835 adaptive detection algorithms, adaptive receivers for spread-spectrum system and, 100-105 adaptive differential PCM satellite communications and, 880 speech coding/synthesis and, 2343, 2354, 2355, 2372, 2382, 2820–22, 2822 adaptive digital filters, 687, 699-700, 701-702 adaptive equalizers, 79-94 adaptation algorithm in, 82 baseband equivalent channel in, 80, 81 baseband transmission system, 79-80, 79 Beneveniste-Goursat algorithm in, 92 blind (see also blind equalizers), 79, 82, 91-93, 286-298 channel estimator in, 90 classification of, and algorithms used in, 81-82, 81 constant-modulus algorithm in, 92 decision-directed mode in, 82 decision-feedback, 81, 87-89, 88, 89 delayed decision feedback sequence estimation, 81 digital video broadcasting and, 91 discrete Fourier transform and, 86 distortion and noise in, 79 fast linear equalization using periodic test signals in, 86 fast startup equalization in, 82 filters in, 81-82, 89 finite impulse response transversal filter in, 81 high frequency communications and, 955 intersymbol interference and, 79-81, 87, 1161-62 lattice filter in, 81 least mean square algorithm in, 83-84, 85, 88, 90 least squares, 81-82, 84-85 linear feedback shift register in, 85 linear, 82-87, 82 M algorithm in, 81 map symbol-by-map symbol, 89, 89 maximum a posteriori detector in, 79, 81, 89 maximum length signal generator in, 85 maximum likelihood sequence estimation in, 79, 81, 90 maximum likelihood, 89-91 mean squared error, 81, 83-84, 84, 86, 88 microwave and, 2569-70 midamble in, 90 minimax criterion in, 81, 82-83 multiple input/multiple output, 93 noniterative algorithms in, 82 Nyquist theorem and, 86 orthogonal frequency division multiplexing and, 93 passband transmission system, 80, 80 pseudonoise in, 85 recursive least squares algorithm in, 82, 84-85, 85, 90 reduced state sequence estimation, 81 reference signals and, 85-86 Sato algorithm in, 92 signal to noise ratio in, 88 spacing in, symbol-spaced vs. fractionally spaced, 86–87, **87** startup equalization in, 81 stop-and-go algorithm in, 92 system model for, 79-80, 79, 80 tap leakage algorithm in, 87 training mode in, 82 trellis-coded modulation and, 91, 91 tropospheric scatter communications and, 2700-02 Viterbi algorithm in, 81, 90 whitened matched filter in, 89 zero forcing, 81, 82-83, 88 adaptive filters acoustic echo cancellation and, 6 chann/in channel modeling, estimation, tracking, 413 packet rate adaptive mobile receivers and, 1892-1900, 1893 adaptive integral method, antenna modeling and, 173 adaptive loading, orthogonal frequency division multiplexing and, 1878, 1878

adaptive multirate coder, speech coding/synthesis and, 2828 adaptive postfiltering, speech coding/synthesis and, 2346 adaptive receivers, blind multiuser detection and, 304-306, 305 adaptive receivers for spread-spectrum systems, 95-112 access methods for, 95-96, 96 adaptive detection algorithms for, 100-105 additive white Gaussian noise and, 97, 102 advanced mobile phone service and, 95 auxiliary vector filter algorithm in, 104 bandwidth in, 95 binary signaling and, 109 blind adaptive algorithms in, 100 blocking matrices in, 105 chip duration in, 96 co channel interference and, 96 coding division multiple access and, 96, 96 conventional (matched filter) receiver in, 97-98, 106 correlation matrix estimation in, 107-108 cost or objective function in, 99-100 cross-spectral reduced-rank method for, 104 decision-directed vs. decision feedback methods in, 105, 105 decorrelating detector in, 98 differentia phase shift keying in, 107 differential least-squares algorithm in, 107 differential MMSE and, 107 direct sequence CDMA in, 96, 97, 101, 104, 107 direct vs. indirect receivers in, 99, 99 effective spreading coding for, 106 equal-gain combining in, 106 error sequences in, tentative vs. current vs. prediction, 101 filters in, 103-104 forgetting factor in, 101 frequency division duplex and, 96 frequency division multiple access and, 95-96, 96 gain vector in, 101 Global System for Mobile and, 96 individually optimal detector in, 99 interference and, 95 interferer multiplication in, 108 intersymbol interference and, 103 jointly optimal detector in, 98-99 k-means clustering algorithm in, 103 learning algorithms in, 103 least mean squares algorithm in, 100-101 linear minimum probability of error receivers in, 101-102 linear receivers in, 98 mininum mean squared error receiver in, 98, 101, 102, 103, 106-109 MOE algorithm in, 109 mulituser systems in, 95, 95 multipath interference and, 95, 105-108 multiple access interference and, 97-98, 101-102, 103 multiple access systems in, 95-96, 96 multiple data rates and, 108-109, 109 multipoint to multipoint communications in, 95 multistage Wiener filters in, 104 near-far problems and, 98 optimal receivers in, 102-103 point to multipoint communications in, 95 principal components method for, 104 processing gain and, 96 projection matrix selection for, 104 radial basis function in, 102 RAKE receivers in, 108 recursive least mean squares algorithm in, 101 reduced-rank adaptive MMSE filtering in, 103-104 reduced-rank detection in, 104-105 signal model for, 97-99 statistical multiplexing in, 95 time division duplex and, 96 time division multiple access and, 95-96, 96 training signals in, 100 wireless communicaitons and, need for, 96-97 adaptive round robin and earliest available time scheduling, 1556, 1559

adaptive routing, routing and wavelength assignment in WDM and, 2102 adaptive transform coding, 2837 adaptive vector quantization, 2128 adaptive virtual queue, 1661 add drop multiplexers, 482, 1634, 1637, 2493-94, 2494 additive increase multiplicative decrease, 1630, 1662, 2439 additive white Gaussian noise adaptive receivers for spread-spectrum system and, 97, 102 bit interleaved coded modulation and, 275-286 blind equalizers and, 287, 288 blind multiuser detection and, 298 cable modems and, 327, 328, 331 cellular communications channels and, 393 chann/in channel modeling, estimation, tracking, 410 chaotic systems and, 424 chirp modulation and, 442, 445-447 coding division multiple access and, 459, 462 continuous phase modulation and, 589, 2180 convolutional coding and, 599, 601-602, 605 demodulation and, 7, 1335 digital phase modulation and, 709 discrete multitone and, 745 diversity and, 732, 733 fading and, 786-787 finite geometry coding and, 805 image and video coding and, 1034 impulsive noise and, 2402-2420 information theory and, 1114 low density parity check coding and, 1312, 1313 low density parity check coding and, 658 magnetic recording systems and, 2253, 2259, 2261, 2262, **2264** magnetic storage and, 4, 1332 matched filters and, 1336–1337 minimum shift keying and, 1468 multicarrier CDMA and, 1522, 1526 multidimensional coding and, 1542-43, 1545-47, 1546 optical communications systems and, 1487 orthogonal frequency division multiplexing and, 1874 packet rate adaptive mobile receivers and, 1886, 1887, 1888, 1901 permutation coding and, 1954 phase shift keying and, 712 power control and, 1983 product coding and, 2012 pulse amplitude modulation and, 2024-25, 2030 pulse position modulation and, 2037 quadrature amplitude modulation and, 2046, **2046** rate distortion theory and, 2069-80 satellite communications and, 1251 sequential decoding of convolutional coding and, 2143, 2144, 2155, 2156 serially concatenated coding for CPM and, 2180 sigma delta converters and, 2237-38, 2238 soft output decoding algorithms and, 2295, 2296 space-time coding and, 2325, 2326 synchronization and, 2473-85 terrestrial digital TV and, 2547 trellis coded modulation and, 2623-24, 2624, 2629 trellis coding and, 2636, 2648 turbo coding and, 2703 ultrawideband radio and, 2756, 2757 unequal error protection coding and, 2765 Viterbi algorithm and, 2816-17 wireless multiuser communications systems and, 1605-06, 1607 wireless transceivers, multi-antenna and, 1580 address of record, session initiation protocol (SIP) and, 2199 address resolution protocol, 548-549 addressing, 547-549 Ethernet and, 1503 local area networks and, 1282 mobility portals and, 2194 paging and registration in, 1914 adjacent channel interference, 530, 1876 admission control, 112-128, 1906 algorithms for, 116-118 ALOHA and, 123

ATM and, 114, 116 ATM and, 205, 1656 bandwidth brokers in DiffServ and, 115 broadband ISDN and, 112-114, 113 burst level, 120 burst switching and, 122 bursty transmission and, 121 call level, 120 call setup and release and, 113-114 cdma2000 and, 366-367 channel borrowing and, 125, 125, 125 channels allocation and, 121, 122-123, 122 circuit switching and, 122 coding division multiple access and, 120, 121, 126 common packet channel switching and, 123 congestion control and, 112 connection admission control, 205 COPS protocol and, 116 data networks and, 116-117 deterministic approach to, 117 Differentiated Services and, 114, 115 distributed, 118 endpoint, 118 enhanced data rate for global evolution and, 126 flow control and, 1625, 1655-56 frequency division multiple access and, 120 general packet radio system and, 126 Global System for Mobile and, 126 guard channels and, 124, 124, 124 handoffs and, 120 handoffs and, 123-126, 124, 123 higher data rate and, 126 hybrid schemes for, 123 Integrated Services (IntServ) and, 114-115, 115 International Mobile Telecommunications 2000 and, 126 Internet and, 114, 115-116 measurement-basedadmission control in, 1656 measurement based, 118 model based, 117 multimedia networks and, 1563-64 multiple link approach to, 117-118 multiprotocol label switching and, 116 network to network interface and, 113 neural networks and, 1681 North American TDMA and, 126 overview of, 112, 120-121 packet level, 120 packet reservation multiple access and, 123 packet switching and, 122 policy-based (policy enforcement point; policy decision point), 115, 118 power control and, 121-122 powerline communications and, 2004 private network to network interface and, 113-114 quality of service and, 112, 114, 115-117, 116, 120, 121, 122, 126 queuing priority and, 124-125, 125 radio resource management and, 2093-94 resource-based algorithms and, 112, 118 resource reservation protocol and, 114-115, 115, 116 satellite onboard processing and, 482 service level agreements and, 115 signaling system 7, 113 single link approach to, 117 soft and safe, 2094 statistical multiplexing and, 2428-29, 2428 TCP/IP and, 114 third-generation standards for, 125-126 time division CDMA, 123 time division duplex, 123 time division multiple access and, 120, 123 traffic models for, 117 Universal Mobile Telecommunications Systems and, 120, 126 user to network interface and, 113-114 wideband CDMA and, 126 wired networks, 112-128 Advanced Communications Technology Satellite, 1227, 1228 advanced encryption standard, 606, 608, 610, 1152, 1648 advanced mobile phone service, 95, 1478-1480 cochannel interference and, 455 IMT2000 and, 1095-1108 interference and, 1130-41 IS95 cellular telephone standard and, 347 satellite communications and, 2116 time division multiple access and, 2586 Advanced Research Projects Agency, 267 Advanced Telecommunication Technology Satellite, 483 Advanced Television Systems Committee and, 2549, 2550 advanced video coding in, 1054-55, 1055 aeronautical communications, antennas for mobile communications and, 198-199 affine projection algorithm, acoustic echo cancellation and, 7-8, 8 AFOSR project, 1739 agglomerative methods in quantization and, 2128 aggregated route-based IP switching, 1599 AIR, wireless MPEG 4 videocommunications and, 2978 air interface standard, cdma2000 and, 359-367 airborne/warning and control system, antennas and, 169 airgap matching networks, 1410-11, 1410, 1412 Alamouti scheme, 1584-85, 1611, 1611, 1619 alarm indication signal, ATM and, 207 algebraic CELP, 1304, 1306, 2349, 2355, 2356, 2826-27 algebraic replicas, multidimensional coding and, 1541-42 algebraic vector quantized CELP, 1306 Algorithm A, sequential decoding of convolutional coding and, 2140, 2145-46 all optical network, wavelength division multiplexing and, 2843-45, 2843 All Optical Networking Consortium, 1720 allocation-based protocols, media access control and, 5, 1343, 1344–1346 allowed cell rate, ATM and, 552 ALOHA protocols, 128-132, 268 admission control and, 123 ATM and, 2907-09 Bluetooth and, 315 capture, 130-131, 130 carrier sense multiple access and, 129, 341-344, 341, **342**, 346 cdma2000 and, 366-367 coding division multiple access and, 131 collision resolution algorithms in, 130-131 drift analysis in, 129, 129 finite number of users model for, 129 frequency division multiple access and, 825 frequency division multiplexing and, 130 improving efficiency of, 129 infinite number of users model for, 128-129, 128 media access control and, 1346, 1347, 1552, 1553, 1559 multibase, 131 multichannel and multicopy, 130 nonbinary transmissions in, 131 optical fiber and, 1720 packet rate adaptive mobile receivers and, 1902-03 quantitative analysis of slotted, 128-129 rebroadcasting using, 128 reservation, 130 satellite communications and, 1232, 1253 shallow water acoustic networks and, 2208, 2209 slotted, 128, 341-342, 500-501, 500 spread spectrum and, 131, 132 throughput and, 342 throughput per slot rate in, 128 time constraints in, 131 time division multiplexing and, 130 traffic engineering and, 499-501, 499, 500 unslotted, 128, 342-343, 343 wavelength division multiplexing and, 2842 alpha trackers, in channel modeling, estimation, tracking, 415 alternate mark inversion, 1934 American Mobile Satellite Corporation in, 2112 ammonium dihydrogen phosphate transducers (acoustic), 34 AMOS 8 feed, waveguides and, 1392, 1392

Ampere-Maxwell laws, antennas and, 171 amplified spontaneous emission, optical fiber systems and, 1842–48, 2272 amplifiers community antenna TV and, 512, 517 erbium doped fiber amplifiers (see erbium doped fiber amplifiers) lasers and, 1776-77 optical communications systems and, 1484, 1485, **1486**, 1707, 1709–10, **1710**, 1848 satellite onboard processing and, 477 semiconductor optical amplifiers (see semiconductor optical amplifiers) solitons and, 1767 variable gain, 2250 amplitude distribution in skywaves, 2063 amplitude modulation, 132-141, 679-680, 1477-78, 1825 active antennas and, 49, 50 analog signal and, 132-133 balanced modulator for, 139, 139 cable modems and, 332, 332 carrier signal and, 132, 133 community antenna TV and, 518-519, 519 conventional double sideband, 133, 134-135, 135 double sideband, 133 double sideband suppressed carrier, 133-134, 133, 140. 140 envelope detectors for, 134-135, 135, 139-140, 139 filters and, 134, 135–136 Hilbert transform, Hilbert transform filters in, 135 lowpass filter and, 134, 136 message signal and, 132, 133 millimeter wave antennas and, 1425 mixers and, 139 modulators and demodulators (modems) for, 137-140, 1497 noise and distortion in, 134 optical transceivers and, 1826-30, 1826 overmodulation in, 134 phase coherent (synchronous) demodulator for, 134 power law modulation and, 138, 138 ring modulator for, 139, 139 sidebands in, upper and lower, 133 single sideband, 133, 135–136, **136**, 140, **140** suppressed carrier signal in, 134 switching modulator for, 138-139, 138 vestigial sideband, 133, 136-137, 137, 140 voltage spectrum of, 134 amplitude probability distribution, impulsive noise and, 2402-2420 amplitude shift keying (see also digital phase modulation), 709–719 chirp modulation and, 444 power spectra of digitally modulated signals and, 1988, 1989-91, 1991 pulse amplitude modulation and, 2022-23 signal quality monitoring and, 2273 amplitude/phase predistorter, predistortion/compensation in RF power amplifiers and, 533, 533 AMRIS ad hoc wireless networks and, 2891-92 AMRoute ad hoc wireless networks and, 2892 AMSC satellite communication, 196 analog signal, amplitude modulation and, 132-133, 2106-11 analog to digital converter/conversioin in acoustic modems for underwater communications, 17 - 18cable modems and, 327 channelized photonic, 1964-65, 1964 digital filters and, 686-687 distributed mesh feedback, 1967 electrooptic, 1961-64, 1962, 1963 frequency synthesizers and, 833-835, 834 image and video coding and, 1026-27 magnetic storage and, 1319 modems and, 1495 multibeam phased arrays and, 1520, 1521 optical folding flash type, 1963-64, 1963 oversampling, 1965-68, 1965 photonic, 1960-70, 1961

sampling and, 2106-11, 2106 sigma delta converters and, 2227-47, 2228 software radio and, 2305, 2306, 2308, 2313 speech coding/synthesis and, 2370 in underwater acoustic communications, 43 waveform coding and, 2830 wireless multiuser communications systems and, 1609 analysis by synthesis (AS), 2344-50, 2344, 2823-24 analysis, wavelet, 2846-62 analytical path loss prediction models, radiowave propagation and, 216 angle error, millimeter wave propagation and, 1435 angle modulation methods (see also frequency modulation; phase modulation), 807-825 angle of arrival, 2689-90, 2689, 2963-64 anomalous propagation, microwave and, 2559-60 anomaly detection, 1652 anonymous file transfer protocol, 1152 antenna arrays, 141-169, 180 adaptive, 163 antenna characteristics and indices in, 142-144 applications for, 187 array factor in, 142-169 attenuators in, 166 bandwidth in, 144 Bayliss line source and, 155-157, 156, 157 binomial, 187 binomial linear, 144 bit error rate and, 163 broadside, 145, 187 Butler matrix feed in, 166, 167 Chebyshev, 187 Chebyshev binomial linear, 145, 187 Chebyshev linear, 152, 152, 153-154, 154 Chebyshev, 145-148, 148 circular, 142, 149-151, 150, 151 cochannel interference and, 455 coding division multiple access in, 163 concentric ring circular, 150-151 conformal, 142, 152-153, 152 conical conformal, 152-153 coupling in, 160 current distribution in, 142 cylindrical conformal, 152, 152 cylindrical, 151-152, 152 dichroic, 142 digital beamforming, 142, 163-164-164 directional characteristics in, 141 directivity gain in, 142 distribution, synthesis by, 154 Dolph-Chebyshev linear, 145-146, 147, 187 element patterns and coupling in, 164-166, 165, 166 elements and array types in, 142, 144 endfire, 142, 145, **148**, **149**, 187 far field in, 141 feeds for, 166, 166, 167 finite vs. infinite, 165-166 flatplate slot, 142 Fourier transform and orthogonal method in, 157-158, 158 fractal, 142 frequency division multiple access in, 163 genetic algorithm optimization in, 162-163, 163 geometry of, 141-142, 141 Gram-Schmidt procedure and, 158 half power beamwidth of, 143, 153 Hansen-Woodyard endfire, 145 hemispherical conformal, 152-153 impedance and, 160 index in, optimization of, 160, 161 intelligent, 163 iteration and, modified patterns by, 156-157, 157 Legendre linear, 148 line sources and distributions in, 154 linear broadside, 142 linear, 144-148, 144 magnitude of, 144, 145 method of moments in, 165 microstrip patch, 152, 152, 187, 1371-1377, 1380-1390 multibeam phased arrays in, 1513-21

multiple antenna transceivers for wireless communications and, 1579-90, 1580 null beamwidth in, 143 optimization in, 160-164 optimization in, by index, 160, 161 optimization in, simplex and gradient method for, 161 orthogonal perturbation method in, 159, 159, 160 orthosynthesis in, 158-159, 159 pattern function of elements in, 142 pattern synthesis for, 153-157 phase shifters in, 166 photonic feeds in, 166 planar, 142, 148-149, 149, 150 plane radiation pattern, principle plane in, 142-144, 143 polarization of antennas in, 142 power gain in, 143 Poynting vectors in, 165 quality factor in, 143 radiation efficiency in, 143 radiation intensity in, 142-143 radiation patterns and, 142-144, 143, 160 Riblet linear, 147, 147, 148 root matching, synthesis by, 154 sampling, synthesis by, 154 scanning type communications and, 152, 187 series feed in, 166 shunt feed in, 166 sidelobe level in, 144 signal to noise ratio in, 143 simulated annealing optimization in, 161-161 slotted, 142 smart antennas in, 163 space and time optimization in, 163-164, 163, 164 spatial division multiple access in, 163 spatial processing, 163 spherical conformal, 152-153 spherical coordinates of, 142 supergain in, 160-161 synthesis as optimization problem in, 160, 187 Taylor distribution (Chebyshev error) and, 154-157, 155, 157, 187 Taylor one-parameter distribution in, 155 thinned arrays in, 162–163 3D, 151–152, **152** time division multiple access in, 163 total electric field in, 142 total magnetic field in, 142 uniform linear, 144-145, 146, 153, 153 visible region in, 144 Woodward-Lawson method and orthosynthesis in, 158-159, 159, 187 Yagi-Uda, 161, 161, 187 antenna beam switching, satellite onboard processing and, 478-479 antenna duplexers, surface acoustic wave filters and, 2458-59 antenna index, 160 antenna modeling techniques, 169-180, 182-184 absorbing boundary conditions in, 177 adaptive integral method (AIM) in, 173 Ampere-Maxwell laws in, 171 aperture source modeling in, 183–184, 183 assembly process in, 177 basis functions in, 174, 175 boundary element/boundary integral methods in, 176 coupling integrals in, 174 differential approach to, 170 Duffy's transform in, 174 electric field integral equation in, 173 entire-domain basis functions in, 174 expansion coefficients in, 174 expansion coefficients in, 176-177 Faraday law and, 171 fast algorithms in, 173 fast Fourier transform in, 173 fast multipole method in, 173 field equivalence principle in, 183-184 finite element (FE) method for, 170, 176-177, 176 finite-element boundary integral methods in, 170, 177 frequency domain based, 169, 170

antenna modeling techniques (continued) gain in, 170 Galerkin method in, 174, 177 Gauss elimination in, 173 Green function in, 172, 174, 176 Helmholtz equations in, 171 Huygen's principle in, 183-184 hybrid techniques for, 170, 177 integral approach to, 170, 172–176 Kirchhoff's current law and, 175 lower-upper decomposition in, 173 magnetic field in, 171 Maxwell's equations and, 169, 170-172, 176 method of moments in, 173, 174, 175 Nystrom's method in, 173 perfect electric conductor in, 183-184 permeability in, 171 permittivity in, 171 point matching or colocation method in, 174 quadrilateral grid in, for onboard auto antennas, 171 Rao-Wilton-Glisson basis functions in, 176 rooftop basis functions in, 176 slot antennas and, 170 source current and spheical boundary in, 171, 171 source modeling and, 182-184 Strat-Chu equation in, 172 subdomain basis functions in, 174, 175-176 surface equivalence principle in, 172, 172 surface modeling in, 175-176, 176 thin-wire theory in, 175 time domain based, 169 time domain/integral equation methods in, 169 weighting functions in, 175 weighting in, 176 wire modeling in, 174–175, 174, 175, 182–183, 183 antennas, 179-188 absorbing boundary conditions in, 177 active (see also active antennas), 47-68 adaptive, 180, 184, 192, 192 adaptive arrays (see adaptive antenna arrays) aeronautical communications, 198-199 Ampere-Maxwell laws in, 171 aperture efficiency in, 186 aperture source modeling in, 183-184, 183 aperture-type, 180, 184 arrays of (see antenna arrays) atrmospheric refraction and, 210-211, 210 bandwidth in, 144, 169, 180 beamforming, 191-192, 192, 480 beamwidth and, 185 beverage, 1259 Bluetooth, 169 broadpattern pattern, 180 built in, 194-195 categories of, 188 cavity backed cross slot, 197-198 cellular communications channels and, 393 cellular telephone, 169, 183, 189, 189 chip, 195–196 cochannel interference and, 454-455 conformal, 169 cordless telephone, 189 corner reflector, 191, 191 coupling in, 160 crossed dipole, 199 crossed drooping dipole, 197, 197 crossed slot, 199 dead zones and, 215 diffraction in, 213-215, 214, 215, 215-216 dipole (see dipoles) directional, 198 directivity gain in, 142 directivity in, 180, 185, 186, 196 divergence factors in, 211 dual beam, 191, 194 dual frequency, 191, 194 early researches into, 179 earth reflection and, 209–210, **209** effective area of, 180, 186 electric field in, 180 electrical equivalents in, 180

elements of, 180 elements, in arrays, 144 fan, 180 far field (Fraunhofer) region in, 181-182, 182 Faraday law and, 171 feeds for, 166, 166, 167 field equivalence principle in, 183-184 field regions in, 181-182, 182 figures of merit for, 180, 184-186 four-third's earth radius concept in, 210 free space propagation equations in, 208-209, 209 frequencies and, 179, 180, 190, 191, 192, 193 frequency independent, 180 Fresnel reflection coefficient for, 209 Fresnel zones and, 214 Friis equation in, 2015 gain in, 169, 185-186, 190, 192-193, 196 gain to system noise in, 196 Global System for Mobile and, 194 Green function in, 172, 174, 176 ground reflection point, 209–210 half power beamwidth of, 143, 153, 185 half wave, 193 helical and spiral, 180, 183, 193-194, 193, 935-946, 936-945 Helmholtz equations in, 171 high frequency (HF) communications and, 951 high gain, 169 horn, 179, 180, 184, 187, 1006-17, 1006, 1392, 1392 Huygen's principle in, 183-184, 214 impedance and, 160, 169, 177, 177, 180, 184, 186 index of, 160 indoor propagation models and, 2015 isotropic radiator, 185 Kirchhoff's current law and, 175 klystron, 179 leaky wave, 1235-47 lens type, 180 linear, 1257-60 lobes in, 184 log periodic, 169, 187 long wire, 180, 188 loop, 183, 1290-99 loss in, 186 magnetic field in, 171, 180 magnetron, 179 main (major) lobe in, 184 Maxwell's equations and, 169, 170-172, 176, 179, 180-181 meander, 193, 194, 194 method of moments in, 173, 174, 175 microstrip, 180, 184, 193 microstrip/microstrip patch (see microstrip/microstrip patch antennas) microwave, 179, 180, 2567 military, 169 millimeter wave, 1423-33 mobile communication (see also antennas for mobile communications), 169, 188 modeling (see antenna modeling techniques) monopole, 183, 193, 193 multibeam phased arrays and, 1513–21, **14** multiple antenna transceivers for wireless communications, 1579-90, 1580 multiple input/multiple output systems and, 1450-56, 1450 near grazing incidence in, 210 null beamwidth in, 143 Nystrom's method in, 173 omnidirectional, 197-198 paging system, 183 parabolic, 1920–28, 1920 parameters of, 184-186 patch, 169, 180, 193 path loss and, 216-217, 1936-44 pattern beamwidth and, 169 perfect electric conductor in, 183-184 planar inverted F, 193, 195, 195 plane radiation pattern, principle plane in, 142-144, 143 polarization diversity, 191

polarization efficiency in, 186 polarization in, 142, 180, 186, 196 power density in, 185, 186 power gain in, 143 propagation factor or path gain factor in, 209 proximity effects and, 189 quad helical, 198 quadrifilar helical, 197, 197, 199 quality factor in, 143, 199 quarter wave, 193 radiating near field (Fresnel) region in, 181-182, 182 radiation density in, 185 radiation efficiency in, 143, 184–186 radiation intensity in, 142–143, 185 radiation patterns and, 142-144, 143, 160, 175, 180-181, 181, 184, 190, 192, 193 radiator, 180, 199 Rayleigh criterion and, 212-213, 213 reactive near field region in, 181-182, 182 reflector, 169, 179, 180, 184, 187, 2080-88 resistance in (radiation and loss), 184 rhomboid, 180 roughness factors (specular effects) and, 211-213, 212, 213 rubber duck and, 193 satellite, 169, 189, 196-199, 477-479, 877-878 scattering and, 215 short backfire, 198, 198 sidelobe level in, 144, 184 sidelobes in, 169 signal to noise ratio in, 143 slot, 169, 180 smart, 180, 184, 187, 191 Snell's law and, 210 space-time coding and, 2324-32, 2324 spatiotemporal signal processing and, 2333-40, 2333 standing wave, 1257 Strat-Chu equation in, 172 supergain in, 160-161 surface equivalence principle in, 172, 172 surface roughness (specular effects) and, 211-213, 212, 213 switched beam, 191-192 television and FM broadcasting, 180, 187, 2517-36 theory of, 180-182 Thevenin equivalent circuits and, 184, 185 thin-wire theory in, 175 transverse electromagnetic waves and, 182 waveguide aperture, 179 waveguides and (see waveguides) wire modeling in, 174-175, 174, 175 wireless communications and, 169, 179-180, 183, 184, 190 Yagi-Uda, 169, 187 antennas for mobile communications, 188-200 adaptive, 192, 192 aeronautical communications, 198-199 bandwidth in. 190 base station, 190-192 beam tilting in, 190, 190 beamforming, 191-192, 192 built in, 194-195 carrier to noise level ratio in, 190 cavity backed cross slot, 197-198 cellular telephone, 189, 189 chip type built in, 193 chip, 195-196 cordless telephone, 189 corner reflector, 191, 191 crossed dipole, 199 crossed drooping dipole, 197, 197 crossed slot, 199 design of, 189–193 directional, 198 directivity in, for satellite communications, 196 diversity reception in, 190 dual beam, 191, 194 dual frequency, 191, 194 fading in, 190 frequencies for, 190, 191, 192, 193 frequency domain duplexing, 190

antennas for mobile communications (continued) gain in, 190, 192-193, 196 gain to system noise in, 196 generations of development in, 189 Global Positioning System, 198 half wave, 193 helical, 193-194, 193 L band, 196 mean effective gain in, 192-193 meander patch antenna, 193, 194, 194 microstrip patch, 193, 199, 197, 197 mobile station, 192-196 monopole, 193, 193 multipath propagation and, 190 Navigation System with Time and Ranging in, 198 omnidirectional, 197-198 passive intermodulation effects and, 191 planar inverted F, 193, 195, 195 polarization diversity, 191 polarization in, 196 proximity effects and, 189 quad helical, 198 quadrifilar helical, 197, 197 quality factor in, 199 radiation patterns in, 190, 192, 193 radiators, 199 Rayleigh distribution and, 190 requirements of, 188-189 satellite, 189, 196-199 selection criteria for, 189 short backfire, 198, 198 smart, 191 space division multiple access in, 191 switched beam, 191-192 terrestrial (land-mobile) systems and, 189-196 wireless, 190 anticipation, constrained coding techniques for data storage and, 575 antiguiding parameters, chirp modulation and, 447 antipodal neural networks and, 1676 antipodal signaling, minimum shift keying and, 1457 antireflection coatings, lasers and, 1779 AntiSniff, 1646 Apache servers, wireless application protocol and, 2900 aperture leaky wave antennas (LWA) and, 1238 parabolic and reflector antennas and, efficiency in, 1923-24, 2080-81 waveguides and, 1406-09, 1406, 1408, 1409, 1419-20 aperture coupled microstrip/microstrip patch antennas and, 1362-63, 1363, 1368-1370, 1371 aperture efficiency, antennas, 186 aperture error, cable modems and, 328, 329 aperture source modeling, antennas, 183-184, 183 aperture type antennas, 180, 184 apodization, surface acoustic wave filters and, 2450-52 apogee and perigee in orbit, 1248 application layer OSI reference model, 540 packet switched networks and, 1911 streaming video and, 2436 TCP/IP model, 541 application level data units, medium access control and, 1553 application level security, 1155 application programming interfaces (API), 1651, 2311 archival systems, magnetic storage and, 1319 area coverage, cell planning in wireless networks and, 374 area spectral efficiency, cochannel interference and, 454 areal density, hard disk drives and, 1321, 1322 ARIB, Bluetooth and, 309 arithmetic coding, 636-638, 1032 ARPA Packet Radio Program, 268 ARPANET, 267-268, 2653 array antennas (see antenna arrays) array factor, active antennas and, 62-63, 62, 142-169 array gain, multiple input/multiple output systems and, 1450, **1450**, 1451 arrayed waveguide grating, 1752-54, 1753, 1786-90, 1787, 1788

active optical cross connects and, 1790-96 cyclic port shifting in, 1787 cyclic wavelength shifting in, 1787 free spectral range in, 1787 passive optical switches and, 1789-90, 1789 periodicity in, 1787 port compatibility in, 1788-89, 1788 reciprocity in, 1787 self-blocking ports in, 1788-89 signal quality monitoring and, 2271-72 symmetry in, 1787 arrayed waveguide grating router, 1723, 1724, 1731, 1731 arrays (see antenna arrays) arrival process, traffic engineering and, 486-491 arrival statistics, traffic engineering and, 486 arrivals, traffic engineering and, 489 articulation, in speech coding/synthesis and, 2360, 2364-65 articulation index, in speech coding/synthesis and, 2362, 2363-68 artificial neural networks, 1675-83, 2378-79 ascending node in orbit, 1248 ASCII, 546 assembly process, antenna modeling and, 177 association, disassociation, reassociation, wireless communications, wireless LAN and, 1287 associative memory, in neural networks and, 1677 assured forwarding, DiffServ, 270-271, 669, 670-673, 670, 671 assured rate, DiffServ, 675 astra return channel system, 2120 Astrolink, 484, 2112 asychronous balanced mode, 546 asymmetric ciphers, cryptography and, 1152 asymmetric digital subscriber line, 272, 1500 architecture design for, 1572, 1573 bit error rate and, 1573-75 broadband wireless access and, 317 channel gain to noise ratio in, 1573 cost minimization in, 1573-75 discrete multitone and, 746-747, 747 error resilient entropy coding in, 1576 image data over, 1576 multicarrier modulation in, 1572 multimedia over digital subscriber line and, 1570, 1571-72, 1571 parallel transmission in, 1572, 1574-75 peak signal to noise ratio in, 1576-77, 1576, 1577 quadrature amplitude modulation and, 1576 quality of service and, 1573, 1575-76 satellite communications and, 2121 serial transmission in, 1572, 1574, 1575 signal to noise ratio in, 1573 subchannel to layer assignment in, 1574-75 system optimization in, 1572-76 time slot assignment in, 1574, 1575 unequal error protection coding and, 2767 video over, 1576 asymmetric key/public key encryption, 606, 607, 611-612 asymptotic coding gain, 2628 asymptotic multiuser efficiency, 461-465 asynchronous connectionless link, Bluetooth and, 313, 315, **315** asynchronous response mode, 546 asynchronous transfer mode, 549-553, 550 admission control and, 114, 116, 205, 1656 alarm indication signal in, 207 allowed cell rate in, 552 ALOHA protocols and, 2907-09 asymmetric digital subscriber line and, 272 ATM adaptation layer in, 264, 266 ATM block transfer in, 267 ATM Forum and, 266, 272 ATM layer in, 200, 206-207, 264 available bit rate in, 206, 267, 551, 1658, 1663 Banyan networks in, 202–203, **202** Batcher-banyan networks in, 203 BISUP protocol and, 204 broadband and, 2655, 2659-61, 2660

broadband ISDN and, 204, 264-267, 271-272 buffering input and output in, 201, 203-204, 203 burst tolerance in, 551 bus matrix switch in, 203-204 carrier sense multiple access and, 345 cell delay variation in, 551 cell delay variation tolerance in, 266 cell forwarding in, 200 cell loss priority in, 200, 206, 550, 1659 cell loss ratio in, 266, 550 cell tax in, 264-265, 273 cell time and switching in, 201 cell transfer delay in, 551 cells in, 550, 1658 closed-loop rate control in, 206 code division multiple access, 2907-09 community antenna TV and digital video in, 524 congestion avoidance and control in, 551-552 connection admission control, 205 connection control or control plane in, 200, 204 connection establishment protocols in, 552-553 connection oriented nature of, 265, 550 connection setup in, 2911-12, 2912 constant bit rate in, 206, 266, 551-553, 1658, 1663 contention in, 201 continuity checking in, 207 control plane in, 264 controlled cell transfer in, 267 credit-based control schemes in, 551 crossbar switch in, 202-203, 202 cyclic redundancy check in, 264 data over cable service interface specifications and, 272 dense wavelength division multiplexing and, 273 efficient reservation virtual circuit in, 552-553, 552 error detection and correction in, 2908-09 Ethernet and, 1512 explicit cell rate in, 552 explicit forward congestion indication in, 200, 206 failure and fault detection/recovery in, 1633-34 fast resource management in, 552 fault management in, 206-207 fault tolerance and, 1633, 1635 flow control and, 550, 1625, 1654, 1656 generic cell rate algorithm in, 201, 205-206, 205, 266, 267, 1656, 1659 guaranteed frame rate in, 1658 handover in, 2912-14, 2912 header error control in, 200, 201, 550 headers in, 200, 550 HiperLAN and, 2909 input and output port processing in, 201 IP telephony and, 1181 IP traffic and, 273 layered architecture of, 200-201 loopback calls in, 207 management informatioin base for, 200 management plane in, 264 MASCARA protocol and, 2908 maximum burst size in, 117, 266, 551, 1656, 1658 maximum cell transfer delay in, 266 medium access control and, 2907-09 minimum cell rate in, 266, 552 multimedia cable network system and, 272 multimedia networks and, 1567 multiprotocol label switching and, 1594, 1598-99 multistage interconnection networks in, 202, 202 network network interface in, 264, 265 neural networks and, 1681 non real time variable bit rate in, 551, 1658 nonreal time VBR in, 206, 267 NxN crossbar switching in, 201-203, 202 open shortest path first and, 204 open vs. closed loop control in, 551 operations and maintenance function in, 200-201, 206 optical fiber and, 1719, 2615, 2619-20 output buffered switches in, 201 output contention in, 201 packet scheduling in, 206 packet switched networks and, 1909

asynchronous transfer mode (continued) packets in, 200, 264 peak cell rate in, 266, 551, 552, 1656, 1658 peak to peak cell delay variation in, 266 performance management in, 207 permanent virtual circuit in, 204, 265-266 private network node interface and, 204, 205, 1635 Q.2931 standard in, 204 quality of service and, 204, 205, 207, 266, 272, 273, 550, 552, 1658 queues in, 201 rate-based control schemes in, 551-552 real time variable bit rate in, 206, 267, 551, 1658 reference model for, 264, 264 reliability and, 1633-34, 1635 remote defect indicator in, 207 resource management in, 552 routing in, 205 satellite communications and, 2113, 2115, 2120 scalability of switches in, 201 security and, 1154 selective cell discarding in, 206 service specific connection oriented protocol and, 2619 - 20shared memory/shared medium for, in switching, 201-202, 202 signaling ATM adaptation layer protocol and, 204 signaling in, 204, 204, 2909-14 simple network management protocol in, 200 SONET and, 201, 273 space division switching in, 202-203 statistical multiplexing and, 2420-32 sustainable cell rate in, 117, 266, 551, 1656, 1658 switched virtual circuits (SVC) in, 265-266 switching in, 200-207, 200, 272-273 synchronous digital hierarchy and, 201 TCP/IP and, 264, 273 traffic contracts in, 205 traffic engineering and, 273 traffic management in, 205-206, 266 traffic modeling and, 1672 traffic shaping in, 551-552 unequal error protection coding and, 2767 unspecified bit rate in, 206, 267, 551, 1658 usage parameter control in, 205, 206 user network interface in, 264, 265, 272, 1657 user plane in, 264 variable bit rate in, 267, 2420 virtual channel connections in, 1658 virtual channel identifier in, 201, 206, 264, 270, 549, 550 virtual channels in, 200, 205, 550 virtual circuit deflection protocol in, 553 virtual circuits in, 207, 264, 270, 550, 1635 virtual path identifer in, 200, 201, 206, 264, 270, 549, 550 virtual paths in, 200, 205, 264, 273, 550 virtual private networks and, 273 wavelength division multiplexing and, 273, 2845, 2864 wireless and, 2906-15, 2907 wireless LANs and, 2681 asynchronous transmission, 545-546, 545, 1495, 1808 AT&T. 262, 370 ATM adaptation layer, 264, 266 ATM block transfer, 267 ATM Forum, 266, 272, 273 ATM layer, 200, 206-207 ATM-88x modem, 18, 18 atmospheric effects, 1434-43, 2558-60, 2559 atmospheric noise, 949, 2061, 2067, 2405-12 atmospheric particulate effects, millimeter wave propagation and, 1439-43 atmospheric radiowave propagation, 208-217, 2059-69 amplitude distribution in skywaves and, 2063 atmospheric noise and, 2061 atmospheric refraction and, 210-211, 210 attenuation and, 215-216 Cairo curves and sky waves in, 2061 cell planning in wireless networks and, 375-377 chaotic systems and, 428-431

critical frequency and, 2065 data bank of propagation paths for, 2063 daytime measurement of, 2064 dead zones and, 215 diffraction in, 213-215, 214, 215, 215-216, 2013, 2018 diurnal variations in, 2063, 2065 divergence factors in, 211 earth reflection and, 209–210, **209** extremely low frequency in, 758-780 fading and, 781-802, 2065 FCC clear channel skywave curve in, 2061-62 field strength in, 2066 field strength in, predicted vs. measured, 2064-65 flat earth approximation and, 209 four-third's earth radius concept in, 210 free space propagation equations in, 208-209, 209 Fresnel reflection coefficient for, 209 Fresnel zones and, 214 ground conductivity and, 2060 ground reflection point in, 209-210 ground wave propagation in, 208, 2059-60 high- and low-latitude curve and skywaves in, 2061 high frequency in, 946-958, 2059-60 high latitude anomalies in, 2065-66 Huyghen's principle and, 214 indoor propagation models for, 216-217, 2012-21 IONCAP software for, 2066 ionospheric propagation and, 208, 2059, 2060, 2065 Kirke method and groundwave propagation in, 2060 latitude and, 2064 low frequency, 2059-69 magnetic coordinates and, 2061 magnetic field activity and, 2063-64 mateial properties and, 2013-14 maximum usable frequency and, 2065, 2066 medium frequency, 2059-60 millimeter wave propagation in, 1433-50 Millington method and groundwave propagation in, 2060 model for, 2064-65 multipath and, 2065 near grazing incidence in, 210 noise and, 2061 north south curve and skywaves in, 2061 outdoor propagation models for, 216 path accuracy in, 2064 path loss and, 216-217, 1939-41 pathlength in, 2064 propagation factor or path gain factor in, 209 Rayleigh criterion and, 212-213, 213 reflection and, 2013, 2018, 2065 region 2 skywave method in, 2061-62 roughness factors (specular effects) and, 211-213, 212, 213 satellite, 208 scattering and, 215, 2013, 2018-19, 2692-2704 seasonal variations in, 2063 seawater effects on, 2064 skip distance and, 2065 sky wave propagation and, 208, 2061-65 Snell's law and, 210 solar cycles and, 2060-61 solar flares and, 2066 space wave propagation in, 208 sporadic E and, 2065 spread F in, 2065 Stokke method and groundwave propagation in, 2060 sudden ionospheric disturbance and, 2066 sunspot activity and, 2061, 2063, 2065 surface roughness (specular effects) and, 211-213, 212, 213 tropical anomalies in, 2065 troposphere and, 208, 2059 tropospheric scatter and, 215, 2692-2704 Udaltsov-Shlyuger skywave method in, 2062, 2064 VOACAP software for, 2066 Wang skywave method and, 2062-63, 2064 atmospheric refractive turbulence (see also scintillations), 1861-63, 1861 atrmospheric refraction and, 210-211, 210

attachment unit interface, 1506 attack signature decision, 1652 attenuation, 30, 30 acoustic echo cancellation and, 4-5 cellular telephony and, 1479 community antenna TV and, 517-518, 517 free space optics and, 1855-57, 1856 local multipoint distribution services and, 1273, 1276–77, **1276** microwave and, 2560 millimeter wave propagation and, 1270-72, 1271, 1443-45, 1444, 1445-48, 1446, 1447 millimeter wave propagation and, rain and precipitation, 1440-45, 1440, 1441 optical fiber systems and, 435, 439, 1708, 1708, 1709, 1710-11, 1714, 1843, 1844, 1982-83 power control and, 1982-83 powerline communications and, 2000, 2001 radiowave propagation and, 215-216 waveguides and, 1405, 1405 wavelength division multiplexing and, 2869 attenuators, antenna arrays and, 166 audibility, in speech coding/synthesis and, 2363-64 audio H.324 standard for, 918-929, 919, 918 orthogonal frequency division multiplexing and, 1867, 1878 audio coding, terrestrial digital TV and, 2552-53 audits, security, 1650 AUSSAT satellite communication, 196 authentication, 1151, 1152, 1647, 1649 Bluetooth and, 316 cdma2000 and, 364 central authority in, 613-614 cryptography and, 606, 607, 611, 613-614 Diffie Hellman coding in, 614 Fiat Shamir identification protocol in, 614 general packet radio service and, 875 global system for mobile and, 906 manipulation detection coding in, 613 message authentication coding in, 613 public key infrastructure in, 614 Schnorr identification protocol and, 614 trusted authority in, 613-614 trusted third party in, 613-614 virtual private networks and, 2810 wireless communications, wireless LAN and, 1287, 1288 zero knowledge in, 614 authentication coding, 218-224 authentication with arbitration model for, 222 bucket hashing and, 221-222 Cartesian product construction in, 223, 223 construction of, 221-222 hash functions and, 221-222 impersonation attack vs., 219, 222 model system for, 218-219, 219 nontrusting parties and, 222-224 perfect, equitable, and nonperfect, 220 probability of deception and, 219, 223 projective plane construction in, 220-221 properties of, theorems for, 219-221 Shannon's theory and, 218 Simmons' bounds and, 219-220 source message in, 219, 222 square root bound in, 220 substitution attack vs., 219, 222 systematic (Cartesian), 220, 221 validation process in, 219 authentication header, virtual private networks and, 2811, 2811 authentication with arbitration model, 222 authoritative name servers, 548 authorization, 1647 autocorrelation fading and, 783 feedback shift registers and, 795, 798, 799 free space optics and, 1862 impulsive noise and, 2402-2420 linear predictive coding and, LS, 1262 orthogonal frequency division multiplexing and, 1945 autocorrelation (continued) packet rate adaptive mobile receivers and, 1889-90 peak to average power ratio and, 1945 polyphase sequences and, 1975 power spectra of digitally modulated signals and, 1990 pulse amplitude modulation and, 2023-24 signature sequence for CDMA and, 2276-85 ternary sequences and, 2541-42 traffic modeling and, 1667, 1669, 1669 automatic bias control, optical modulators and, 1746 automatic gain control in acoustic modems for underwater communications, 17 - 19cable modems and, 327 microwave and, 2567 pulse amplitude modulation and, 2026 satellite onboard processing and, 477 automatic link estabilshment, 951-952, 2313-14 automatic protection switching, SONET and, 2494-95, 2495[°] automatic repeat request, 224-231, 545, 1632 acknowledgments in, 226 block coding and, 225 Bluetooth and, 313, 314 cdma2000 and, 364 check bytes and, 225 crossover probability in, 230, 230 cyclic redundancy check and, 225-226 detection and correction coding in, 230-231 efficiency and reliability of, 228-230 feedback (return) channel and, 224 forward error correction and, 230-231 frame error rate in, 228-230, 229 frame structure and, 225-226, 225 go back N, 226-227, 227, 229-230, 545 Hamming coding and, 225, 229-230 high frequency communications and, 952 hybrid method for, 225, 230-231 incremental redundancy principle in, 231 linear coding and, 225 multimedia over digital subscriber line and, 1571 negative acknowledgment in, 226 OSI reference model and, 225-226 parity bits and, 225 performance analysis of, 228-230 powerline communications and, 2002, 2004 protocols for, 226-228 radio resource management and, 2093 satellite communications and, 879, 1223, 1229-31, 1230, 1231 selective reject ARQ, 545 selective repeat, 228, 228, 229-230 shallow water acoustic networks and, 2207-08, 2210 - 12sliding window protocol in, 227 stop and wait, 226, 226, 229-230, 545 throughput and, 226, 228-230 trellis coded modulation and, 225, 2635 undetected errors and, 225 weighting, Hamming weight, weight enumerator) in 225-226 wireless multiuser communications systems and, 1612 - 13automatic speech recognition (see also speech coding/synthesis), 2373-79, 2382, 2383-90, 2385 acoustic model for, 2385 artificial neural networks in, 2378-79 cepstrum in, 2373, 2386 comb filtering in, 2378 continuous speech recognition in, 2377 current state of, 2389-90 difficulties of, 2383-84 dynamic time warping in, 2373 feature extraction stage in, 2384, 2386-87 hidden Markov models and, 2373-80, 2385 history and development of, 2384-89, 2384 interactive voice response systems and, 2384 language models for, 2376-77, 2385, 2388-89 linear prediction coders in, 2373 mel scale frequency cepstral coefficients in, 2373, 2382

noise in, 2378 recognition stage in, 2384 relative spectral method in, 2378 segmentation in, 2377-78 speaker verification systems and, 2379-80 stochastic approach to, 2373-74 timing problems in, 2373 training stage in, 2384, 2387 Viterbi algorithm and, 2818 automorphism, Golay coding and, 889-890, 889 automotive collision avoidance systems, 503 autonomous ocean sampling network, 24, 25, 2211-12 autonomous systems, 549, 1153 IP networks and, 269 packet switched networks and, 1913 autonomous underwater vehicles (AUV), 24, 36, 2206, 2211-12 autoregressive process, 412 autoregressive moving average, 412, 2293 autoregressive processes, traffic modeling and, 1666, 1668 auxiliary vector filters, 1890–96, 1891, 1893 auxiliary vector filter algorithm, 104 available bit rate, 123, 206, 267, 551, 1658, 1663 avalanche photodiode detectors, 1002, 1834, 1857, 1962 avalanche shot noise, optical fiber systems and, 1843 average interference power, signature sequence for CDMA and, 2283 average magnitude difference function, speech coding/synthesis and, 2350 average matched filter, chirp modulation and, 446 axis of the deep sound channel, in underwater acoustic communications, 38 back propagation algorithm, neural networks and, 1678 back to back user agents, session initiation protocol and, 2198 background limited infrared performance, 1858 backhauling, 1636, 1636 backoff, Ethernet and, 1281, 1346 backsearch limiting, sequential decoding of convolutional coding and, 2153-54 backup schemes, 1319, 1634-35 backward error correction, 545 Bahl-Cocke-Jelinek-Raviv decoding, 556, 561-564, 564 low density parity check coding and, 1316 soft output decoding algorithms and, 2295, 2297, 2299–2301, **2299** space-time coding and, 2328 tailbiting convolutional coding and, 2515 Bahl-Jelinek algorithm, 2738 balanced driving receivers, 1827 balanced incomplete block design, 658, 659-661, 659, 1316 balanced modulator, amplitude modulation and, 139, 139 bandgap lasers and, 1777, 1778 bandpass filters/modulators, 1478 impulsive noise and, 2415-16 optical signal regeneration and, 1764 pulse amplitude modulation and, 2022 quadrature amplitude modulation and, 2044-45, 2045 random processes, sampling of, 2286-87 sampling and, 2108-2111, 2109 sampling of, 2286 sigma delta converters and, 2238-40, 2239, 2240 signal quality monitoring and, 2272 bandwidth, 37-38, 549, 2653 in acoustic modems for underwater communications, 15, 16 acoustic telemetry in, 22 adaptive receivers for spread-spectrum system and, 95 antenna, 144, 169, 180 antennas for mobile communications and, 190, 192 batching in, 234-235, 234 Bluetooth and, 310-311 cable modems and, 324-325 coding division multiple access and, 459 compression and, 631 continuous phase modulation and, 589-590, 2180

digital phase modulation and, 709 digital phase modulation, Nyquist criterion, 710 free space optics and, 1849-1850 IS95 cellular telephone standard and, 350 local multipoint distribution service and, 318, 1268 magnetic storage and, 1326 microstrip/microstrip patch antennas and, 1360, 1364-1370 millimeter wave antennas and, 1425, 1434 minimum shift keying and, 1457 multibeam phased arrays and, 1518 multimedia networks and, 1562, 1563, 1565, 1568 multiprotocol label switching and, 1598 optical cross connects/switches and, 1784, 1797 optical fiber and, 436, 1719-20, 1732, 1797 optical modulators and, 1744 optical transceivers and, optimization in, 1836-37 packet switched networks and, 1908, 1906-07 partial response signals and, 1929, 1932-33 patching in, 233–234, 234 periodic broadcasting and, 235-236, 236 photodetectors and, 1000 piggybacking in, 232-233, 232 pulse amplitude modulation and, 2023 quadrature amplitude modulation and, 2043, 2045-46, 2046 reduction techniques for (see also bandwidth reduction techniques for video service), 232-237 serially concatenated coding for CPM and, 2180 shallow water acoustic networks and, 2207 software radio and, 2315-17 space-time coding and, 2327 speech coding/synthesis and, 2363, 2364, 2365-67, 2365 statistical multiplexing and, 2428-29 traffic modeling and, 1666 trellis coding and, 2636-37 ultrawideband radio and, 2761-62 underw/in underwater acoustic communications, 36 waveguides and, 1390 wavelength division multiplexing and, 2865 wireless infrared communications and, 2927 wireless LANs and, 2678 wireless multiuser communications systems and, 1603, 1604 bandwidth brokers, 115, 1568 bandwidth reduction techniques for video services, 232-237 Banyan networks, ATM and, 202-203, 202 Barker coding polyphase sequences and, 1980-1981 in underwater acoustic communications, 43 wireless LANs and, 2942 base station antennas for mobile communications and, 190-192 cell planning in wireless networks and, 375, 376 cellular communications channels and, 393 local multipoint distribution service and, 318-319 wireless multiuser communications systems and, 1612-13, 1613 base station controller, 866-876, 905-17 base station diversity, IS95 cellular telephone standard and, 355-356 base station location powerline communications and, 1999 satellite communications and, 2117 wireless multiuser communications systems and, 1603 base station subsystem, 866-876, 905-17 base transceiver station, 866-876, 905-17 baseband communications, 79, 79 adaptive equalizers and, 79-80, 79 in channel modeling, estimation, tracking, 398-401 discrete multitone and, 741-742 pulse amplitude modulation and, 2022 pulse position modulation and, 2034-35 synchronization and, 2479 tapped delay line equalizers and, 1690, 1690 baseband equivalent channel, adaptive equalizers and, 80, 81 baseband filters, orthogonal frequency division multiplexing and, 1872

baseband PAR, orthogonal frequency division multiplexing and, 1945 baseline privacy, 324, 335 basic service set, wireless communications, wireless LAN and, 1285 basis functions, antenna modeling and, 174, 175 Batcher-banyan networks, ATM and, 203 batching, bandwidth reduction and, 234-235, 234 batwing antennas, television and FM broadcasting, 2517-36 baud vs. fractional rate, 286, 1496 baud/symbol loop, cable modems and, 328-329 Baum-Welch algorithm, hidden Markov models and, 961-962 Bayesian estimation, in channel modeling, estimation, tracking, 398 Bayes estimation of random parameter, 2, 1340 Bayliss line source, in antenna arrays and, 155-157, 156, 157 BCH coding (see also BCH coding, binary; cyclic coding), 616-630 Berlekamp decoding algorithm for, 624-625 bounds in, 621-624 cyclotomic cosets in, 622 decoding, 622-626 elementary symmetric functions in, 623 Euclid's algorithm and, 617 generating functions in, 623 Massey-Berlekamp decoding algorithm for, 625-626 narrow sense, 622 power sum symmetric functions in, 623 primitive, 622 trellis coding and, 2640 in underwater acoustic communications, 43 BCH coding, binary (see also BCH coding, nonbinary), 237 - 253Berlekamp iterative decoding algorithm in, 247, 250-251 block coding and, 243-252 channel measurement decoding in, 247 Chien search decoding in, 249-250, 250 codeword, codeword polynomial and, 244 cyclic block coding and, 243-244 cyclic coding and, 243-252 cyclic redundancy check and, 245 decoding of, 247-252 design distance in, 245 elementary symmetric functions in, 248 encoding circuit with shift register for, 246-247, 246 erasure filling decoding in, 247, 251-252 error control and, 238 error locators in, 247-248, 250, 623 error trapping decoding in, 247, 251 extending, 246 extension fields and, 240-241 finite fields and, 238-239 forced erasure decoding in, 252 Golay coding and, 245-247, 251 Hamming coding and, 245, 247, 247 irreducible polynomials and, 241-242 Kasami decoding in, 247, 251 logarithm tables for, 239 minimal polynomial properties in, 242-243 minimum functions in, 242 multiple error detection and correction in, 244 nonprimitive elements in, 239 order of element, order of field in, 239 Peterson's direct solution method for, 248-250 polynomial properties defined on finite fields in, 242 polynomials and, 239-240 prime fields in, 239 prime power fields and, 240 primitive element in, 239 primitive polynomials and, 240-241 primitive vs. nonprimitive type, 244-252 reciprocal roots in, 250 Reed-Solomon, 238 shortening in, 246 soft decision decoding in, 251-252 syndrome equations for decoding in, 247-248 t error correcting coding and, 244

vectors and, 239-240 vectors of field elements and polynomials defined on finite fields in, 239-240 Wagner coding and, 252 BCH coding, nonbinary (see also BCH coding, binary; Reed-Solomon coding), 253-262 bounded distance decoding in, 254 Chien search decoding and, 256-257, 260 connection polynomial in, 257-258 decoding in, 254-261 encoding in, 254, 254 erasure filling decoding in, 259-261 error magnitudes or error values in, 254 feedback shift register and FSR synthesis in, 257-259 Fourier transforms and, 261 generalized minimum distance decoding in, 261 locator fields in, 253 Massey-Berlekamp decoding algorithm for, 257-259, 260 maximum coding and, 254 maximum distance separable coding and, 254 modified syndromes for decoding in, 259-260 Newton's identities and, 255, 257, 623, 625 Peterson's direct solution method for, 255-257, 260, 617 polynomials and, 254 primitive vs. nonprimitive types, 253 soft decision decoding in, 261 symbol fields in, 253 syndrome equations for decoding in, 255, 623 t error correcting coding and, 253 beam deviation factors, parabolic and reflector antennas and, 1926, 1927 beam patterns, satellite communications and, 877-878 beam propagation method, optical modulators and, 1745 beam scanning, parabolic and reflector antennas and, 2084 - 86beam steering millimeter wave antennas and, 1431, 1432 multibeam phased arrays and, 1519, 1520 beam switching, satellite onboard processing and, 478-479, 478 beamforming antennas, 191-192, 192, 2963 adaptive antenna arrays and, 163-164, 187 multibeam phased arrays and, 1517, 1517, 1518, 1520-21, 1520 satellite onboard processing and, 480 software radio and, 2307 spatiotemporal signal processing and, 2333-40, 2333 beamforming network, waveguides and, 1393 beamshaping free space optics and, 1851 wireless transceivers, multi-antenna and, 1579 beamwidth antennas and, 169, 185 array antennas, 187 leaky wave antennas and, 1239 local multipoint distribution services and, 1273 multibeam phased arrays and, 1517 parabolic and reflector antennas and, 1922-23, 1925-26 beat noise/distortion in, 514, 1836 behavior aggregate, DiffServ and, 270 Bell System, 262 Bell, Alexander Graham, 262, 1849 Bellman-Ford algorithm, 2208 bending, millimeter wave propagation and, 1435 bending radius, optical fiber and, 438-439, 438 Beneveniste-Goursat algorithm, equalizers and, 92 Berlekamp decoding algorithm BCH coding, binary, and, 247, 250-251 BCH/ in BCH (nonbinary) and Reed-Solomon coding, 624-625 cyclic coding and, 624-625 Berlekamp-Massey algorithm, 790, 797-798 Bernoulli sources, rate distortion theory and, 2073 Bessel functions, optical modulators and, 1742 best effort forwarding, IP networks and, 269 best effort networks, packet switched networks and, 1910

best effort service, radio resource management and, 2094–95

beverage antennas, 1259 bias maximum likelihood estimation and, 1339 neural networks and, 1676 bidirectional path switched ring, SONET and, 2495-96 bidirectional self-healing ring in, 750, 751 Big Leo satellite communications, 1251 bilateral public peering, 268 billing systems, mobility portals and, 2194 binary bipolar with n zero substitution, 1934 binary convolutional coder, CATV, 526-527, 526 binary frequency shift keying, 16 minimum shift keying and, 1457 satellite communications and, 1225, 1225 binary orthogonal keying, chirp modulation and, 441, 444, 445 binary PAM, serially concatenated coding and, 2165 binary phase shift keying, 371, 408, 410, 710-711, 2179 acoustic telemetry in, 23 blind multiuser detection and, 298 cdma2000 and, 362 IS95 cellular telephone standard and, 354 orthogonal frequency division multiplexing and, 1945, 1948 peak to average power ratio and, 1945, 1948 polyphase sequences and, 1976 predistortion/compensation in RF power amplifiers and, 530, 531 satellite communications and, 1225, 1225, 1230 serially concatenated coding and, 2165 serially concatenated coding for CPM and, 2180 soft output decoding algorithms and, 2296 trellis coding and, 2638-53 turbo coding and, 2704-16 turbo trellis coded modulation and, 2738-53 ultrawideband radio and, 2755-62 wireless multiuser communications systems and, 1610, 1614 binary signaling, adaptive receivers for spread-spectrum system and, 109 binary symmetric channel low density parity check coding and, 1315 rate distortion theory and, 2073 sequential decoding of convolutional coding and, 2143, 2144, 2146 binomial linear antenna arrays, 144 biphase coding, constrained coding techniques for data storage and, 576 birefringence, optical, 1492, 1711 BISUP protocol, ATM and, 204 bit allocation, 646, 1044-46, 1045 bit error probability (BEP) power control and, 1983 quadrature amplitude modulation and, 2043, 2049, 2050-52 2052 trellis coded modulation and, 2629 bit error rate, 2179 antenna arrays and, 163 asymmetric DSL and multimedia transmission in, 1573 - 75bit interleaved coded modulation and, 275 cable modems and, 326, 328 chirp modulation and, 444, 445-447, 445, 446 coding division multiple access and, 458, 459-460 concatenated convolutional coding and, 559-560, 560 continuous phase modulation and, 2180, 2181-89, 2183 convolutional coding and, 599, 602-605 digital phase modulation and, 709 diversity and, 732, 733 fading and, 787 failure and fault detection/recovery and, 1633-34 free space optics and, 1859, 1859, 1862, 1865 hard disk drives and, 1320 holographic memory/optical storage and, 2138 impulsive noise and, 2402-2420 low density parity check coding and, 1309, 1309, 1316 magnetic recording systems and, 2266, 2266 measuring, 2572-75, 2573, 2574, 2575 microwave and, 2565-67, 2566

3006 INDEX

bit error rate (continued) multidimensional coding and, 1545-48, 1548 optical cross connects/switches and, 1785 optical fiber and, 2614 optical fiber systems and, 1841, 1846-47, 1971-73, 1971 packet rate adaptive mobile receivers and, 1886, 1887, 1892, 1898, 1902 phase shift keying and, 713-15 powerline communications and, 2004 predistortion/compensation in RF power amplifiers and, 530, 535, 536 pulse position modulation and, 2039, 2039 Reed-Solomon coding for magnetic recording channels and, 473 satellite communications and, 881, 1224, 1225, 1227, 1230, 2120 semianalytical MC technique in, 2293-94 sequential decoding of convolutional coding and, 2156, **2156** serially concatenated coding for CPM and, 2180-89, 2183 signal quality monitoring and, 2269, 2270 simulation and, 2293-94 software radio and, 2314 space-time coding and, 2330 terrestrial digital TV and, 2547 trellis coded modulation and, 2629 trellis coding and, 2637, 2638, 2639 turbo coding and, 2704-16 turbo trellis coded modulation and, 2738, 2747-49 in underwater acoustic communications, 37 unequal error protection coding and, 2763-69 Universal Mobile Telecommunications System and, 387 wavelength division multiplexing and, 655 wireless and, 2923, **2924** wireless infrared communications and, 2927 bit flipping, low density parity check coding and, 1311-12. 1312 bit interleaved coded modulation, 275-286, 275, 277, 275 additive white Gaussian noise and, 275-286 bit error rate and, 275 channel state information and, 280 constellation labeling in, 279 decoder /encoder for, 279-283, 279 free distance and, 279 Gray labeling and, 281, 282, 284-285 Hamming distance and, 278–279, 281, 282 interleaving in, 276, 276 iterative decoding in, 284 maximum likelihood detector in, 277, 283 metric generators for, 283 multipath fading and, 276, 278 orthogonal frequency division multiplexing and, 278 performance characteristics of, 284-285, 285 quadrature amplitude modulation and, 281 quadrature phase shift keyed and, 279 Rayleigh fading channel in, 278, 280, 281, 283, 285 shift register encoder for, 278-279 signal to noise ratio and, 277, 278, 285 time diversity and, 276 trellis coded modulation (TCM) and, 276-286, 276 trellis decoding in, 279-280, 279, 282-283 turbo coding and, 285 Ungerboeck set partitioning and, 276, 276, 280-281 Viterbi algorithm and, 280 wireless communications and, 276 bit interleaved parity, signal quality monitoring and, 2269 bit interval, modulation, 1335 bit loading, discrete multitone and, 745 bit oriented transmission, 546 bit pipes, 539 bit rates modems and, 1496 speech coding/synthesis and, 2341 synchronous digital hierarchy and, 2496-97 bit stuffing, 547 Blahut algorithm, rate distortion theory and, 2075

BLAST architecture, spatiotemporal signal processing and, 2333 blind adaptive multiuser detectors, 464 blind adaptive receivers for spread-spectrum system and, 100^{-1} blind carrier recovery, 2054-56 blind channel estimation, 402, 404-407 blind clock recovery, 2056-57, 2058 blind equalization (see also blind multiuser detection), 82, 91-93, 286-298 adaptive equalizers and, 79 additive white Gaussian noise and, 287, 288 baud rate vs. Nyquist rate sampling in, 287 baud vs. fractional rate, 287-288, 288 carrierless amplitude and phase signals and, 292 channel estimation in, 292-296 channel models for, 287-288, 288 combined channel and symbol estimation in, 289-291 commercial applications for, 296-297 constant modulus algorithms and extensions in, 292 cross-relation approach to SIMO channel estimation in, 294-295 cumulant matching in, 293-294 decision directed algorithms in, 291 decision feedback equalizer in, 289, 290, 292 digital signal processors and, 296 digital subscriber line and, 287, 296 direct equalization and symbol estimation in, 291-292 distortion and, 286 equation error in, 293 expectation maximization algorithm in, 290 filtering matrix in, 290-291 filters in, 288-289 finite impulse response and, 287, 292 fitting error in, 293 fractionally spaced equalizer and, 288-289, 289 Godard algorithms and, 292 hidden Markov model in, 290-291 high order sequence criteria and, 291-292, 297 intersymbol interference and, 286, 291 iterative least squares with enumeration, 292 likelihood function in, 289 maximum likelihood detector in, 289-291, 289 maximum likelihood sequence estimation, 297 mean cost function in, 291 minimum mean square error, 292 multimodulus algorithm in, 292 multistep linear prediction in, 295-296 neural networks and, 1680 noise subspace approach to SIMO channel estimation in, 295 optical fiber systems and, 287, 296 oversampling in, 287, 288–289 probability density function in, 289 quadrature amplitude modulation and, 292, 296 sampling in, 286, 287 Sato algorithm and, 291 Shalvi-Weinstein algorithm in, 292 signal to noise ratio and, 297 single input/multiple output, 292, 294-296 single input/single output model of, 287-288, 288, 291.293 spatiotemporal signal processing and, 2336-38 structures in, 288 systems models for, 287-289 time representation sequence in, 288 training sequences in, 286, 287 wireless communications and, 296-297 blind multiuser detection (see also blind equalizers), 298-307 adaptive implementations for, 300 adaptive receiver structure for, 304-306, 305 additive white Gaussian noise in, 298 batch processing method for, 300 binary phase shift keying and, 298 channel estimation using, 303-304 coding division multiple access and, 298-307 direct matrix inversion in, 298, 300, 306 direct methods of, 299-301

direct sequence spread spectrum and, 298 filtering in, 299 group blind type, 306 intersymbol interference and, 303 least mean square algorithm in, 300-301 mean square error and, 303 minimum mean square error and, 298-307 minimum output energy detector in, 301 multipath channels and, 302–306 multiple access interference and, 299 NAHJ algorithm in, 301, 302, 306 PASTd algorithm in, 301 signal to interference plus noise ratio in, 302, 306 simulation example for, subspace, 301-302 singular value decomposition in, 301 smoothing factors in, 303, 304 subspace approach to, 298, 301-302 blind techniques, spatiotemporal signal processing and, 2333 blindness, in microstrip antennas, 1371, 1376, 1388 block ciphers, 607-609, 607 block coding, 2179 automatic repeat request and, 225 BCH coding, binary, and, 243-252 constrained coding techniques for data storage and, 576-579 image and video coding and, 1038-39 image processing and, 1076 interleaving and, 1141-51, 1142-1149 magnetic recording systems and, 2257 multiple input/multiple output systems and, 1455 product coding and, 2007 satellite communications and, 1229-30, 1230 space-time coding and, 2326-27, 2326, 2329-30 spatiotemporal signal processing and, 2333 threshold coding and, 2583-84 trellis coded modulation and, 2622 block error rate convolutional coding and, 602-605 cyclic coding and, 616-630 Reed-Solomon coding for magnetic recording channels and, 473 block fading channels, wireless multiuser communications systems and, 1605 block missynchronization detection, Reed-Solomon coding for magnetic recording channels and, 471-472 block processing, acoustic echo cancellation and, 12 blocked calls, 1906 blocked calls delayed, traffic engineering and, 495-497 blocked calls held, traffic engineering and, 497-499 blocking, traffic engineering and, 486-487, 486 blocking matrices, adaptive receivers for spread-spectrum system and, 105 blocking probability, traffic engineering and, 498 blue violet lasers, optical memories and, 1739 Bluestein-Gulyaev waves. surface acoustic wave filters and, 2441 Bluetooth, 307-317, 1106, 1289, 2391, 2677 ad hoc communications systems and, 308, 309 ALOHA protocol and, 315 antennas and, 169 applications for, 307-308 asynchronous connectionless link in, 313, 315 automatic repeat request, 313, 314 bandwidth for, 310-311 baseband layer for, 311 coding division multiple access and, 310 collision avoidance in, 315 connection setup process in, 311-312 cyclic redundancy check and, 312 direct sequence CDMA and, 310 embedded radio systems and, 309 error correction in, 313 forward error correction in, 313 frequency division multiple access and, 309 frequency hop spread spectrum and, 309-310 frequency hopped CDMA and, 310, 316 Gaussian frequency shift keying, 310-311, 508 header error coding and, 312 hop selection mechanism in, 312-313

Bluetooth (continued) industrial scientific medical band in, 309, 316 intelligent transportation systems and, 502, 506, 508-510, **508**, **509**, **510** interference and, 309 linear feedback shift registers in, for security, 316 link manager in, 314 link manager protocol (LMP) and, 310, 314 logical link control and adaptation protocol and, 310, 314 low-power modes for, 313-314 MAC addresses and, 312, 315 master slave configuration in, 314-316 networking with, 314-316 packet-based communications and, 312-313, 312 pairing in, for security, 316 park mode in, 314 personal area networks and, 2682, 2683-84 physical links in, 313 piconets in, 314-315, 508 protocol stack for, 310, 310 regulation of, 309 RF layer of protocol stack in, 310-311 scan, page, and inquiry modes in, 311-312 scatternets in, 315-316 security (authentication, encryption) in, 316 sharing spectrum in, 309 spectrum for, 308-310 spread spectrum and, 2400 synchronous connection oriented link in, 315, 315 time division multiple access and, 309-310 time division multiplexing in, 315 time divison duplexing in, 310 unlicensed radio band used in, 308-309 wireless multiuser communications systems and, 1602 Blum Blum Shub random number generator, 615 Blu-Ray Disc, constrained coding techniques for data storage and, 579 BodyLAN, 2681, 2682 Bolt Beranek and Newman, 267, 268 bootstrap effect, serially concatenated coding and, 2166 border gateway multicast protocol, 1535, 1536 border gateway protocol, 269, 549, 550, 1153, 1535, 2809 border gateway protocol 4, 1597 border gateways, general packet radio service and, 867 Bose-Chaudhuri-Hocquenghem (see BCH coding) bottom up processing, in speech coding/synthesis and, 2363 bound, BCH coding, 621-624 boundary conditions, loop antennas and, matching in, 1294-95 boundary element/boundary integral methods, antenna modeling and, 176 boundary routers, DiffServ, 668-669 bounded delay encodable coding, 578 bounded distance decoding, in BCH (nonbinary) and Reed-Solomon coding, 254 bounds, maximum likelihood estimation and, 1, 1339 Box-Mueller method for random number generation and, 2292 Bragg condition, 1756 Bragg gratings (see also diffraction gratings), 1723, 1723, 1727–28, 1727, 1728 bandgap in, 1728 coupled mode theory and, 1728-29, 1729 distributed Bragg reflector lasers and, 1780-81, 1780 erbium doped fiber amplifiers and, 1728 optical add drop multiplexers and, 1727 optical couplers and, 1697-1700 optical fiber and, 1709 optical multiplexing and demultiplexing and, 1749 overcoupling in, 1729 BRAN group, broadband wireless access and, 318-322 branch metrics, trellis coding and, 2647 BRASS high frequency communications and, 948 breathers, solitons and, 1766 Brewster angle in millimeter wave propagation and, 1438 bridges, Ethernet and, 1505 brightness theorem, for active antennas and, 65

Brillouin scattering, 1491, 1684, 1712 broadband communications, 2653-77 asynchronous transfer mode (ATM) and, 2655, 2659-61, 2660 cable modems and, 2668-71 community antenna TV and, 2668-71 data over cable service interface specification and, 2670-2671. 2670 data rates in, 2654-55 delay in, 2655 digital subscriber line and, 2655, 2666-73, 2670 direct video broadcast and, 2671-73, 2672 enterprise networking and, 2656-66 Ethernet and, 2655 fixed wireless, 2671 frame relay and, 2658-59, 2659 future of, 2673-75 Gigabit Ethernet and, 2655, 2656-58, 2657 global network model for, 2655-56, 2655, 2656 global system for mobile and, 2656 Internet protocol and, 2662 IP networks and, 2661-64 IPv4 and IPv6 in, 2663 jitter in, 2655 local multipoint distribution services, 2655, 2671 mobile wireless, 2673 multichannel multipoint distribution services in, 2655, 2671 multiprotocol label switching and, 2655, 2674-75, 2674 packet loss in, 2655 powerline communications and, 1997 reliability in, 2655 residential access to, 2666-73 satellite and, 2112–13, **2113**, 2115, 2655, 2656, 2664–66, **2665**, **2666**, 2671–73 services and applications for, 2654-55 speech coding/synthesis and, 2362 standards for, 2673 TCP/IP and, 2661-63 throughput in, 2655 transmission control protocol and, 2661-62, 2662 user datagram protocol and, 2662 virtual private networks and, 2663-64, 2664 wide area networks and, 2663-64 wireless access (see broadband wireless access) wireless local loop and, 2957-58 broadband integrated services digital network, 262-275 admission control for, 112-114, 113 asymmetric digital subscriber line, 272 asynchronous transfer mode and, 204, 264-267, 271-272 data over cable service interface specifications and, 272 dense wavelength division multiplexing and, 273 high definition TV and, 263 IP networks and, 267-271 local area networks and, 271 MPEG voice compression and, 263 multimedia cable network system and, 272 multimedia networks and, 1567 problems faced by, 273-274 reference model for, 264, 264 SONET and, 273 statistical multiplexing and, 2420-32 virtual private networks and, 273 wavelength division multiplexing and, 273 wide area networks and, 271-272 World Wide Web and, 271-274 broadband radio access network, 318-322, 2958 broadband wireless access, 317-323 BRAN group and, 318-322, 2958 digital audio/video broadcasting and, 318-321 discrete Fourier transform and, 321 frequencies for, 317, 320-322 HiperAccess group for, 319-320 HiperLAN and, 320-321 HiperMAN and, 320 interference and, 318-319 local multipoint distribution service and, 317, 318, 322, 1268-79

microwave multipoint distribution service and, 317 minimum mean square error equalization and, 321 orthogonal frequency division multiple access and, 320-322 quadrature amplitude modulation and, 319, 320 quadrature phase shift keying in, 319, 320 signal to interference ratio in, 319 signal to noise ratio and, 321 single-carrier transmission in, 321-322, 321 standards for, 318, 319-320 time division multiple access and, 318, 320 time division multiplexing and, 318, 320 broadcast domains, Ethernet and, 1281 broadcast satellite service, 877, 1251 broadcasting caution harmonic, 236 digital audio/video, 319-321 fast, 236 frequency modulation, 823, 823, 824 harmonic, 236 media access control and, 1342-1349 pagoda, 236 periodic, 235-236, 236 permutation-based pyramid, 236 pyramid, 236 quasiharmonic, 236 skyscraper, 236 broadpattern pattern antennas, 180 broadside antenna arrays and, 142, 145 Brownian motion models, traffic modeling and, 1669-70 browsers, 540, 548 brute force attacks, 607-608 BSD UNIX, 268 bubble switches, 1792-93, 1792 bucket credit weighted algorithms, medium access control and, 1556 bucket hashing, 221-222 buckets, sequential decoding of convolutional coding and, stack bucket technique in, 2149 buffer management, flow control, traffic management and, 1654, 1656, 1660-61 buffer overflow, sequential decoding of convolutional coding and, 2159-60 buffering allocation and partitioning in, 1565 ATM and, input and output in, 201, 203-204, 203 flow control, traffic management and multimedia networks and, 1562, 1563, 1565-66 occupancy in, 1565 optical fiber and, 2614 packet dropping in, 1565-66 periodic buffer reuse with thresholding, 234 streaming video and, 2438 transmission control protocol and, 1566 bugs and software errors, in security and, 1645 building database, for indoor propagation models and, 2014-15, 2014 built in antennas for mobile communications and, 194 bulk diffraction gratings, 1723, 1723, 1725–26, 1726 bulk handling, differentiated services and, 675 Buratti construction, low density parity check coding and, 661 burst errors, multidimensional coding and, 1540-41, 1544-48 burst mode, Ethernet and, 1284 burst switching networks (see also optical cross connects/switches), 1801-07, 1802 burst assembly in, 1806-07, 1807 burst control packets in, 1801 control channels and control channel groups in, 1803 core routers in, 1802, 1804, 1804 data burst channels in, 1803 delayed reservation in, 1802 edge routers in, 1802 egress edge router in, 1803, 1803 fiber delay lines and, 1804-06, 1805 ingress edge routers in, 1802, 1803 label switched paths in, 1802 quality of service in, 1804-06 reservation protocols in, 1801-02 routers for, 1802-04

3008 INDEX

burst tolerance in, 551 burst transmission, 1906 admission control and, 121, 122 spatiotemporal signal processing and, 2337-38, 2337 bus architecture, 1503, 1504 bus matrix switch, ATM and, 203-204 bus topologies, optical fiber and, 1716, 1716 busy hour, traffic engineering and, 487–488 busy tone multiple access, 1347 Butler beamformer, 1517, 1518 Butler matrix feed, antenna arrays and, 166, 167 Butterworth filters, 686, 2262 C band, 877, 1251, 2113 C means algorithm in quantization and, 2129 cable modem termination system, 324-325, 325, 330, 335 cable modems, 324-336, 1500-01 additive white Gaussian noise, 327, 328, 331 amplitude modulation and, 332 analog to digital conversion in, 327 aperture error and jitter in, 328, 329 automatic gain control in, 327 bandwidth and, 324-325 baseline privacy for, 324, 335 baud/symbol loop in, 328-329 bit error rates and, 326, 328 broadband and, 2668-71 cable modem termination system in, 324-325, 325, 330, 335 channel capacity in, 326-327, 326 data over cable service interface specification and, 324, 327, 333, 334 decision feedback equalization in, 330 destination address in, 324 digital to analog conversion in, 333-334 dynamic host configuration protocol in, 335 effective number of bits in, 327 encryption in, 324, 335 feedforward equalization in, 330 filtering in, 328-329, 324-325, 333 first in first out buffers in, 331-332 forward error correction and, 327, 332-333 frequencies used by, 324-330, 326 head end and, 324 I effects on, 333 intermediate frequency requirements for, 330-334 intermodulation distortion in, 328, 330 intersymbol interference and, 327, 328 interval usage coding and, 334 layered protocols for, 324 least mean squares algorithm in, 330 low noise amplifier and, 327 media access control and, 324, 334-335 minislot usage information in, 334-335 MPEG-2 compression and, 324, 330 Multimedia Cable Network System and, 324 multipath interference in. 334 nonlinearity in (integral and differential) in, 327-328 NTSC standard requirements and, 326 numerically controlled oscillator in, 328 Nyquist property in, 328, 333 packet ID in, 324 packet structure in, 324 PAL standard requirements and, 326 phase noise in, **331**, 331 programmable gain amplifier and, 327, 334 pseudorandom bit sequence in, 328 quadrature amplitude modulation and, 324, 325-328, 330-334, 331 quadrature direct digital frequency synthesis in, 328, 330, 333 quadrature phase shift keying and, 324, 325-328, 330-334 radio frequency interference and, 332 Reed–Solomon coding and, 330, 332 RF frequency spectra requirements for, 326-330 security in, 335 service ID in, 324, 335 session initiation protocol and, 2197-98 Shannon-Hartley capacity theorem and, 326

signal to noise ratio in, 326, 327-328, 327 square root raised cosine filter in, 328-329 surface acoustic wave and, 327 time division multiple access and, 324, 334-335 trellis coded modulation and, 330 trivial file transfer protocol in, 335 upstream channel descriptor and, 334-335 cable TV (see community antenna TV) Cairo curves and sky waves, 2061 California Department of Transportation, 503 call admission control, 1563-64, 1681 call blocking probability, routing and wavelength assignment in WDM and, 2103 call congestion, traffic engineering and, 491, 491 call processing language, session initiation protocol and, 2203 call setup and release, admission control and, 113-114 call to mobility ratio, paging and registration in, 1917 calls, traffic engineering and, 485 capacitance of active antennas and, 49-50, 55 capacitive impedance, active antennas and, 49 capacity (see also Shannon or channel capacity) radio resource management and, 2090 traffic engineering and, 492, 494 capture ALOHA, 130-131 carbon dioxide lasers, 1853 carrier frequency systems, 1996 carrier recovery, synchronization and, 2052-54, 2472-85 carrier selection, for Bluetooth, 312-313 carrier sense multiple access, 339-347 ACK messages in, 345 ad hoc wireless networks and, 2884-86 advanced mobile phone system, 347 ALOHA protocol and, 129, 341-344, 341, 342, 346 asynchronous transfer mode and, 345 capacity of system using, 348-349 channels in, 339 clear to send in, 346 coding division multiple access and, 347-348 collision occurence in, 343-346, 343 Ethernet and, 345, 1280-81, 1503-05, 1504 exponential backoff in, 345 fiber distributed data interface and, 345 frequency division duplexing and, 347 frequency division multiple access and, 347, 349 hidden terminal problem and, 345-346, 345 IS95 cellular telephone standard and (see also), 347-358 local area network and, 345 media access control and, 346, 1346-1347 multiple access channels and, 339 optical fiber and, 1808 point to point communications and, 339 processing gain and, 348-349 protocols used in, 343-346 pseudorandom noise coding in, 349 quality of service and, 346 request to send in, 346 shallow water acoustic networks and, 2209-10, 2212 spread spectrum and, 347, 348 time division multiple access and, 340-341, 340 token ring networks and, 345 wireless communications, wireless LAN and, 1286 wireless LAN and, 346, 2682, 2945 wireless systems and, 2678 with carrier detection, 343-345, 344, 345, 547 carrier signal, amplitude modulation and, 132, 133 carrier suppressed return to zero, 1825, 1828-29, 1829 carrier to interference ratio admission control and, 121 cell planning in wireless networks and, 379, 379 waveguides and, 1416 carrier to noise level ratio antennas for mobile communications and, 190 community antenna TV and, 515-517, 520 carrierless amplitude and phase, 292, 336-338, 337, 338, 2791, 2801 carriers, chaos as, 423-427 Cartesian (systematic) authentication coding, 220, 221 Cartesian coordinate systems, for active antennas, 61

cascade converters, sigma delta converters and, 2234-35, 2234, 2235, 2236-37 CASE tools, software radio and, 2305 Cassegrain parabolic and reflector antennas and, 1920–21, 1921, 2083–86, 2083, 2084 Cassegrain telescope, 1863, 1864 CATA protocol, in media access control and, 1348 catastrophic encoders, 598, 603-604, 604 Category 3 cable, 1283, 1506, 1508 Cauer filters, 686 caution harmonic broadcasting, 236 cavity backed cross slot antenna, 197-198, 197 cavity models, microstrip/microstrip patch antennas and, 1357 CCITT, BISDN and, 262-263 CCSDS standards for cyclic coding and, 629 CDMA to analog handoff, cdma2000 and, 366 cdma2000, 358–369, **358**, 1483, 2391 access control in, 366-367 acronyms pertaining to, 368-369 air interface standard for, 359-367 ALOHA protocols and, 366-367 authentication and message integrity in, 364 automatic repeat request in, 364 binary phase shift keying and, 362 cdmaOne and, 358-359 cell planning in wireless networks and, 369, 385-386 channel structure in, 367 coding division multiple access and, 358-359 compatibility issues and, 367 control on the traffic channel state in, 366 cyclic redundancy check in, 367 diversity in, 361, 367 forward error control in, 359, 360-363 forward fundamental/supplemental coding channels in, 359-362 forward link channels and, 367 forward link in, 359-362, 361 frequency/spectrum allocation for, 358 global positioning system and, 359 handoffs in, 366-367 high data rate packet transmission in, 368 idle state in, 366 IMT2000 and, 1096-1108 initialization state in, 365-366 interleaving in, 359 International Mobile Telecommunications 2000 and, 358 IP networks and, 359 IS95 cellular telephone standard and, 357 key features of, 367-368 link access control and, 359, 364-365, 365 logical channels in, 363-364 media access control and, 359, 363-365 modulation in. 362 multicarrier structure and flexibility in, 358-359 multidimensional coding and, 1548 multiplexing in, 359, 363 orthogonal time division in, 361, 367 OSI reference model and, 359, 365 overlay with TIA/EIA IS95B and, 367 packet data channel control function in, 359, 363, 364 paging in, 366 physical layer for, 359-363 power control in, 366 power management and, 367 protocol data units and, 364-365 pseudonoise coding in, 362 quadrature phase shift keying and, 362 quality of service and, 359, 363 quasiorthogonal functions in, 362 radio link protocol and, 359 reverse link channels in, 367 reverse link in, 362-363, 363, 364 scrambling in, 362 service access points and, 364 session initiation protocol and, 2198 signaling in, 359 signaling radio burst protocol in, 359, 364 spread spectrum and, 359

cdma2000 (continued) spreading coding and, 367 spreading rates in, 359 standards for, 358, 359-367 synchronous base stations in, 367 system access state in, 366 Third Generation Partnership Project and, 358 time division multiple access and, 358 turbo coding and, 367 upper layer (layer 3) signaling in, 365-366 Walsh functions in, 362 wireless local loop and, 2954 wireless multiuser communications systems and, 1602 cdmaOne, 358-359 CD-R media, 1736-37, 1736 CDROM, 1733-35, 1735 CD-RW media, 1737 Celestri, 212 cell delay variation in, 551 cell delay variation tolerance, 266 cell forwarding, ATM and, 200 cell loss priority, ATM, 200, 206, 550, 1659 cell loss probability, admission control and, 117, 118 cell loss ratio, 118, 266, 550 cell planning in wireless networks, 369-393 area coverage in, 374 automatic site placement in, 374-375 base station location and, 375, 376 best server location in, 379 bit error rate in, 387 carrier to interference ratio in, 379, 379 cdma2000 and, 369, 385-386 cell splitting in, 375, 376 channel assignment problem and, 382-383, 383 clutter parameters in, 376, 376 coding division multiple access and, 372 coding puncture rates in, 386 coverage area and, 372 coverage-based results in, 377 digital elevation models in, 374 digital terrain models in, 372 dynamic mode in, 388, 390-391 enhanced data rate for global evolution and, 369, 383–385, **385** Erlang B blocking in, 379–380, 379, 380 financial cost functions in, 374 first-generation sysetms in, 370 forward error control in, 386, 387 fourth generation systems in, 371-372, 391-392 frequency assignment and optimization in, 382-383 frequency assignment problem and, 382-383 frequency division duplex and, 385-386 general packet radio service and, 369, 383-385, 385 geographic functions in, 374 geographic information system in, 372 global conditions and, 382-383 global system for mobile and, 369-373, 377-383 global vs. local parameters for, 372 grade of service and, 379-380 hierarchical approaches to, 375, 375 high speed circuit switched data and, 383-385 history of wireless networks and, 369-372 IMT2000 and, 369, 386, 392 interference and, 377-380 IP networks and, 392 IS136 and, 383 macrocells in, 376 maximum server for, 377, 378 microcells in, 376 minicells in, 376 mobile station location and, 376, 380 network coverage in, 377, 378 network design in, 388-391, 389 network-specific conditions and, 383 Okumura-Hata model for, 376 omni sites and, 374 optimization in, 372 orthogonal variable spreading factors in, 387 probability approach to, 380-382, 381 propagation modeling and, 375-377

protection ratios in, 379 quality of service and, 372, 379, 379, 387 quasidynamic mode in, 388, 389-390 radio network planning tools in, 372, 376, 377, 377 RAKE receivers and, 387, 388 regular grid layout in, 373-375, 373, 374 second generatioin systems in, 370-371, 377-383 service types and, 386-387 signal to interference ratio in, 386, 387 site location and placement in, 372-375 spectral costs in, 374 static mode analysis in, 388, 389 TDSCDMA and, 369, 385-386 third-generation systems in, 371, 385-391 throughput and, 385, 385, 386 time division duplex and, 385-386 time division multiple access and, 377, 380 traffic computation for, 380 traffic coverage rate in, 374 Universal Mobile Telecommunications Systems and, 384, 385-391 wideband CDMA and, 372, 386 WRC2000 and, 392 cell relay, 550 cell splitting, cell planning in wireless networks and, 375, 376 cell switch routers, 1599 cell tax, ATM and, 264-265, 273 cell time and switching, ATM and, 201 cell transfer delay in, 551 cells, 396-393 ATM and, 550, 1658 cellular telephony and, 1479-80, 1479 local multipoint distribution service and, 318-319, 319 wireless multiuser communications systems and, 1602 cellular communications channels, 393-398 additive white Gaussian noise and, 393 antennas and, 393 base station location and, 393 correlation and, 397 crosstalk and, 397 delay spread and, 394, 395 diffraction and, 393 digital advanced mobile phone system and, 397 Doppler shift, Doppler spread in, 394, 395 fading and, 393, 394 frequency allocation and, 393 global system for mobile and, 397 measurement of linear time variant, 395-396 multipath and, 393, 394 Nakagami m distribution and, 394 Rayleigh fading and, 394 reflection and, 393 Rice fading and, 394 scattering and, 393, 394, 395, 395 shadowing and, 394-395 simulation models for, 396-397 standards for, 397 time variance in, 393, 394-395 underspread and overspread, 395 Universal Mobile Telecommunications System, 397 wide sense stationary fading in, 393, 394 cellular digital packet data, 1347, 1350 Cellular Telecommunications Industry Association, 347 cellular telephony (see also multiuser wireless communication systems), 1478-82, 1602-03, 1602 cellular telephony (see also multiuser wireless communication systems), 1603 adaptive antennas and, 192, 192, 454-455 advanced mobile phone system and, 347, 1478, 1479, 1480 antennas and, 141, 169, 189, 189, 393 area spectral efficiency in, 454 attenuation and, 1479 bandwidth in, 190, 192 base station location and, 393 base station antennas for, 190-192 beam tilting in, 190, 190 beamforming antennas in, 191-192, 192

blind multiuser detection and, 298-307 Bluetooth and, 307, 308 built in antennas for, 194-195 carrier to noise level ratio in, 190 cells in, 1479-80, 1479, 1602 Cellular Telecommunications Industry Association and, 347 channels for, 393-398 chip antennas for, 195-196 cochannel interference and, 1480 cochannel interference in (see cochannel interference in digital cellular TDMA networks) coding division multiple access and, 347-348, 458, 1479, 1480 corner illuminated cells in, 450 corner reflector antenna in, 191, 191 correlation and, 397 crosstalk and, 397 delay spread and, 394, 395 development of, 1478 diffraction and, 393 diversity reception in, 190 Doppler shift, Doppler spread in, 394, 395 dual beam antennas in, 191, 194 dual frequency antennas in, 191, 194 Erlang B blocking, 453, **453**, **454** fading and, 190, 393, 394 frequencies for, 347-348, 347, 393, 449, 1478, 1479 frequency division duplexing and, 190, 347 frequency division multiple access and, 347, 829 frequency reuse and, 191, 347, 448, 449-454, 1479, 1480, 1480 global system for mobile and, 1479, 1480 handoffs in, 1479, 1602 helical antennas in, 193-194, 193 intelligent transportation systems and, 506 IS95 cellular telephone standard and, 347-358 location in, 2959-72 macrocells and, 449, 450, 1940-41 mean effective gain in, 192-193 meander patch antenna in, 193, 194, 194 media access control and, 1342-1349 microcells in, 449, 450, 1941 mobile station antennas in, 192-196 Mobile Station Base Station Compatibility Standard for Dual Mode Wideband ... Cellular, 347 mobile switching center in, 1479 monopole antennas in, 193, 193 multipath and, 190, 393, 394 omnicells in, 450 outage probabilities in, cochannel interference and, 451–452, **452** paging and registration in, 1914-28 passive intermodulation effects and, 191 path loss in, 1936-44 personal communication systems and, 1479 picocells in, 449, 450 planar inverted F antennas in, 193, 195, 195 polarization diversity antennas in, 191 power control in, 1982-88 principles of, 1479-81 public switched telephone network and, 1479 Rayleigh fading and, 190, 394 reflection and, 393 Rice fading and, 394 roaming in, 1287 satellite communications and, 2112 scattering and, 393, 394, 395, 395 second generation, 1479 sectored cells in, 450 sectorization in, 454 shadowing and, 394-395 signal to interference ratio in, 1480-81 smart antennas in, 191 space division multiple access in, 191, 455 spectrum efficiency and, 452-454 spread spectrum and, 347, 348 standards, 1479 surface acoustic wave filters and, 2459-60 switched beam antennas in, 191-192 third-generation, 1479

cellular telephony (see also multiuser wireless communication systems) (continued) time division multiple access in, 1479 U.S. digital cellular systems in, 1479 wide sense stationary fading in, 393, 394 wireless local loop standards and systems in, 2947-59, 2948 CENELEC powerline communications and, 1995, 1996, 1997, 2002 center clipper, acoustic echo cancellation and, 1, 1 central authority, authentication and, 613-614 centralized networks, shallow water acoustic networks and 2208 centralized protocols, media access control and, 5, 1343 centroid condition, 2129, 2597 centum call seconds, traffic engineering and, 488 CEO problem, rate distortion theory and, 2076 cepstrum, automatic speech recognition and, 2373, 2386 CEPT, Bluetooth and, 309 ceramic transducers (acoustic) and, 34, 35 chalcogenide (crystal) glass, in optical fiber, 434 channel allocation, in admission control, 121, 122-123 channel assignment problem, 382-383, 383 channel available time table, 1554 channel bits, constrained coding techniques for data storage and, 573, 575 channel borrowing, admission control and, 125, 125 channel capacity (see Shannon or channel capacity) channel coding, 2179 compression and, 631 convolutional coding and, 598 information theory and, 1113 magnetic storage and, 1331-1333 partial response signals and, 1933 rate distortion theory and, 2069 speech synthesis/coding and, 1299 wireless multiuser communications systems and, 1604 channel coherence bandwidth, 1604 channel coherence time, 1604 channel estimation adaptive equalizers and, 90 expectation maximization algorithm and, 771-772 Golay complementary sequences and, 893 space-time coding and, 2330 channel gain to noise ratio, 1573 channel impulse response, 2091-93, 2327, 2474-85 channel measurement decoding, BCH coding, binary, and, 247 channel modeling and estimation (see also channel tracking in wireless systems), 398-408 baseband model in, 398-401 Bayesian estimation in, 398 blind, 402, 404 chaotic systems and, 427 composite baseband channel in, 399 continuous time model in, 398-399, 399 Cramer-Rao bound in, 402, 403, 405, 405 deterministic maximum likelihood algorithm in, 405, 406 deterministic vs. stochastic models in, 405 discrete-time model in, 399-401 estimators for, 401 hidden Markov model and, 406 intersymbol interference and, 398 least mean squares algorithm in, 398, 404 least squares smoothing algorithm in, 407 maximum likelihood algorithm in, 398, 402-405, 427 moment methods in, 403, 406-407 multipath and, 398 multiple input multiple output model in, 400-401, 400 passband signal in, 399 performance bound and identifiability in, 403, 405 performance measure and performance bound, 402 pilot symbols and, 398, 401, 401, 403, 405 point estimation in, 398 projection algorithm in, 407 recursive least squares algorithm in, 404 semiblind, 402, 404-407 single input multiple output model in, 400, 400, 403-407

stochastic maximum likelihood algorithm in, 405-406 subspace algorithm in, 406-407 training based, 401-402 training mode and, 398, 403 channel modeling and identification, neural networks and, 1679 channel optimized coding, waveform coding and, 2830 channel tracking in wireless systems (see also channel modeling and estimation), 408-421 adaptive filters in, 413 additive white Gaussian noise in, 410 alpha trackers and, 415 autoregressive process in, 412 autoregressive moving average process in, 412 coding division multiple access and, 409 complex sinusoidal model in, 412 data directed, joint, 417-418 data directed, separate, 415-417 decision feedback equalizers and, 416, 417-418, 417 delay spread and, 410 demodulator for narrowband, 411 deterministic vs. stochastic modeling in, 411-412 direct sequence CDMA and, 409, 411 Doppler shift, Doppler spread in, 410, 412, 413 equalization in, 417, 417 exponential filtering in, 415 fading and, 410, 410 filters in, 412-414 FIR filters in, 410 global system for mobile and, 409 infinite impulse response filters and, 415 intersymbol interference and, 410, 411, 417 IS136 and, 409 Jakes model for, 410, 411 Kalman model for, 411-412, 411, 414, 415-416 least mean squares algorithm in, 412, 414-415 linear interpolation in, 414, 414 maximum likelihood sequence estimation in, 417-418 mean square error in, 413, 415 models for, 411-412 moving average filters in, 412, 413 multipath and, 410 narrowband systems and, 409-410 per survivor processing in, 416, 418 phase ambiguity problem in, 416 pilot channels and, 409, 413 pilot symbols and, 409, 409, 413-414, 416 RAKE receivers in, 411 random walk in, 412 Rayleigh fading and, 410 recursive approaches for, 414-416 recursive least squares algorithm in, 414, 415 signal to noise ratio in, 413, 414 spread spectrum and, 409 spreading factor in, 411 time division multiple access and, 409-410 training sequences in, 409 Viterbi pruning and, 418 wideband CDMA and, 409 wideband systems and, 409, 411 Wiener filters in, 412, 413, 414 channelization coding, 1975, 2976 channelized photonic AD conversion, 1964-65, 1964 channels carrier sense multiple access and, 339 cellular communications, 393-398 community antenna TV and, 513-514 diversity and, 729 magnetic recording, coding for, 466-476 mobile radio communications and, 1481 modeling and estimation of (see channel modeling and estimation), 398-408 multiple input/multiple output systems and, 1452-53, 1452 optical memories and, 1733 powerline communications and, 2000-2001 satellite communications and, 1224-29 sequential decoding of convolutional coding and, 2042-45

tracking of (see channel tracking in wireless systems) traffic engineering and, 485 chaos in communications, 421-434 additive white Gaussian noise and, 424 carriers and, 423-427 channel encoding and estimation in, 427 chaos shift keyed modulation in, 422, 424, 424, 425-427. 426 chaotic masking modulation in, 423, 423, 423 chaotic pulse position modulation in, 422, 427-428, 427 chaotic switching modulation and, 422, 423-424, 423 coding division multiple access and, 422, 428, 428, 431 coding function in, 427 differential shift keyed modulation and, 422, 425-427, 426 direct sequence CDMA and, 422, 428, 428 drive-response synchronization and, 422, 422 dynamic feedback modulation and, 422, 423, 423 ensemble-averaged autocorrelation in, 428 Euler algorithms and, 424 fading and, 430-431, 430 fractional Brownian motion process and, 431 frequency modulation DCSK in, 422, 425-427, 426 Gold sequences and, 428 Ito-Stratonovich integrations in, 422 laser communications and, 428-431 Lorenz sequence and, 429-430, 429 low probability of intercept (LPI) in, 428 Lyapunov exponents and, 429-430 modulation and, 422, 423, 423 Monte Carlo simulation and, 422 noise and, 421-422 numerical algorithm and performance evalution of, 424-425, 424 radar and, 428-431 radio propagation effects and, 428-431 Runge-Kutta integration and, 422, 424 signal to noise ratio and, 422, 425, 425 spreading sequences and, 422, 428, 428 stochastic differential equation and, 424, 425 sychronization of chaotic systems and, 422 symbolic dynamic models in, 422 symbolic dynamics and, 427 chaos shift keyed modulation, 422, 424, 424, 425-427, 426 chaotic masking modulation, 423, 423 chaotic pulse position modulation, 422, 427-428, 427 chaotic pulse regenerator, 427–428, 427 chaotic switching modulation, 422, 423-424, 423 character oriented transmission, 546 characterization of optical fiber (see optical fiber) charge coupled devices, holographic memory/optical storage and, 2134-35 charge trapping photodetectors and, 1000 chat, 540 cheapernet, 1283 Chebyshev antenna arrays and, 145-148, 148, 187 Chebyshev binomial linear antenna arrays and, 145, 187 Chebyshev error in antenna arrays and, 154-155, 155, 157, 187 Chebyshev filters, 686 Chebyshev linear antenna arrays and, 152-154, 152 check bits, 1308 check bytes, automatic repeat request and, 225 Chien search, 249–250, 250, 256–257, 260, 470, 617 chip antennas, 195-196 chip duration, adaptive receivers for spread-spectrum system and, 96 chirp, 1743-44, 1978 chirp modulation, 440-448 additive white Gaussian noise and, 442, 445-447 amplitude shift keying and, 444 antiguiding parameters in, 447 average matched filter in, 446 binary orthogonal keying and, 441, 444, 445 bit error rate and, 444, 445-447, 445, 446 coding division multiple access and, 445 compression filters for, 442 compression gain in, 443

chirp modulation (continued) differential quadrature PSK and, 444 direct digital frequency synthesizer in, 447 direct sequence CDMA and, 445 Doppler effect, Doppler spreading in, 442 filters for, 446, 447 frequency hopped CDMA and, 445 Fresnel ripples in, 442 full response continuous phase, 444-446 implementation issues for, 447 intersymbol interference and, 443, 446 laser communications and, 447 linear modulation in, 444 matched filtering in, 442–443, **443** multiple access interference and, 445, 446 nonlinear modulation and, with memory, 444-445 partial response continuous phase, 445, 446 performance analysis in, 445-447 phase shifting keying and, 441, 444 postdetection integrator for, 445 pulse position modulation and, 441, 444 Rayleigh fading and, 446 sidelobe reduction in, 443-444 signal to noise ratio and, 442, 445-447 surface acoustic wave filters in, 441, 447 time and frequency representation of, 441-442, 442 time division multiple access and, 445, 446 up- and downchirp frequency in, 441-442, 442 chirp, laser, 1844 chirped return to zero, optical transceivers and, 1830 chromatic dispersion, 436, 1842, 1844, 1845, 1849, 1507, 1784, 2869 cipher block chaining, 335 cipher feedback, 607 CIRC encoders, cyclic coding and, 627-628, 627 circuit switched networks admission control and, 122 failure and fault detection/recovery in, 1633-34 fault tolerance and, 1632 flow control and, 1625 general packet radio service and, 869 H.324 standard for, 918-929, 919 optical cross connects/switches and, 1800 optical fiber and, 2614-15 packet switched networks and vs., 1906, 1906 reliability and, 1632 satellite communications and, 1253-54 shallow water acoustic networks and, 2208 wireless 371 circuits, in microelectromechanical systems and, 1355 circular antenna arrays and, 142, 149-151, 150, 151 circular recursive systematic convolutional coding, 2709, 2709 circulators, optical fiber and, 1709 citizens band, high frequency communications and, 948 cladding, optical fiber and, 434, 435, 1708, 1708, 1714, classes of IP addresses, 269, 548 classified vector quantization, 2127 classless interdomain routing, 269, 1912-13 clear channel skywave curve, 2061-62 clear to send, 346, 1348 client in streaming video and, 2433, 2434 in wavelength division multiplexing, 650-657, **650**–656 clippers, 2416, 2362 clipping, peak to average power ratio and, 1946-47 clock recovery, synchronization and, 2052, 2460, 2472-85 closed loop control, ATM and, 206, 551, 1986 clustering problems and quantization and, 2128 clustering step in quantization and, 2129 clutter parameters, cell planning in wireless networks and, 376, 376 clutter, radar, 429-430, 429 CNET, 264 coarse WDM, 2862 coarticulation, inspeech coding/synthesis and, 2361 coaxial cable, 50, 50 broadband wireless access and, 317

Ethernet and, 1506-07 local area networks and, 1283 microstrip/microstrip patch antennas and feed, 1361–1362, **1362** cochannel interference adaptive receivers for spread-spectrum system and, 96 cellular telephony and, 1480 power control and, 1982 spatiotemporal signal processing and, 2333 wireless multiuser communications systems and, 1604 cochannel interference in digital cellular TDMA networks, 448-458 adaptive antennas and, 454-455 advanced mobile phone system and, 455 area spectral efficiency in, 454 best- and worst-case scenarios for, 454 cancellation of, in time domain, 455-456 coding division multiple access and, 455 corner illuminated cells in, 450 cumulative distribution function and, 451 direct sequence CDMA and, 455 distribution of, 449-451 diversity and, 456 Erlang B blocking, 453, **453**, **454** fading and, 449 Farley's approximation in, 451 Fenton-Wilkinson approximation in, 450, 451 filters for, 454-455 frequency allocation and, 449 frequency reuse and, 448, 449-454 global system for mobile and, 455 loglikelihood ratio and, 456 macrocells and, 449, 450 maximum likelihood sequence estimation in, 455 microcells in, 449, 450 multiple input multiple output and, 455 multiuser detection and, 455 omnicells in, 450 outage probabilities and, 451-452, 452 picocells in, 449, 450 Schwartz-Yeh approximation in, 450-451 sectored cells in, 450 sectorization in, 454 shadowing and, 449 spatial division multiple access and, 455 spatial filtering in, 454-455 spectrum efficiency and, 452-454 time division multiple access and, 453-454, 453, 455 code division multiple access (see also cdma2000), 371, 458-466, 825, 2391 acoustic telemetry in, 25 adaptive antenna arrays and, 187 adaptive receivers for spread-spectrum system and, 96, **96** additive white Gaussian noise and, 459, 462 admission control and, 120, 121, 126 ALOHA protocol and, 131 antenna arrays and, 163 applications for, 458 asymptotic multiuser efficiency in, 461–465 ATM and, 2907-09 bandwidth and, 459 bit error rate in, 458, 459-460 blind adaptive multiuser detectors in, 464 blind multiuser detection and, 298-307 Bluetooth and, 310 cdma2000 (see cdma2000) cell planning in wireless networks and, 372 cellular telephony and, 1479, 1480 channelization coding in, 2876 chaotic systems and, 422, 428, 428, 431 chirp modulation and, 445 cochannel interference and, 455 coding division multiple access (see also optical synchronous CDMA systems), 1817 correlation in optical fiber systems and, 702-709, 703, 705, 708 decorrelator detectors and, 463-465, 463

community antenna TV and, 517-618

dimensionality, processing gain, and, 458-461 direct sequence CDMA and, 458, 459, 1196, 1886-87, 1894, 1975-82, 1975, 2003, 2090, 2091–93, 2209, 2274, 2283, 2284, 2336 enhanced variable rate coder and, 2827 feedback shift registers and, 789 frequency division multiple access and, 829, 2907-09 frequency encoding, 1816–17, **1817**, **1818** frequency hopping, 458, 2276 Gold sequences as, 900–905, 2281–82, **2281**, **2282** Hadamard coding and, 933-934 Hadamard-Walsh coding in, 2874 high frequency communications and, 956 IMT2000 and, 1095–1108, 2873–74 in channel modeling, estimation, tracking, 409 intelligent transportation systems and, 504-505, 507 interference and, 1116, 1119, 1130-41 interference and, 458 interference cancellation in, 1817-23, 1819-23 intersymbol interference and, 2278, 2283 interuser interference and, 461 IS95 cellular telephone standard and, 347-348 Kasami sequences and, 1219-22, 2282 maximal length sequences in, 2279-81, 2280 media access control and, 1343-1345, 1348 microelectromechanical systems and, 1350 minimum mean square error detector in, 463-464 mobile radio communications and, 1481-82, 1482, 1483 multibeam phased arrays and, 1514 multicarrier direct sequence CDMA, 1521-28 multiple access interference and, 458-466, 1196, 2278, 2283 multistage detector in, 464, 464 multitone CDMA and, 1525 multiuser communication systems and, 461 multiuser detection and interference cancellation in, 462-465 near-far problem in, 458, 461-462 neural networks and, 1680-81 on off keying and, 2731-33 optical orthogonal coding and, 2730-31 optical synchronous, 1808-24 orthogonal frequency division multiplexing and, 1878 orthogonal transmultiplexers and, 1880–85 orthogonality of signals in, 458 packet rate adaptive mobile receivers and, 1894 performance measures for, 461 polyphase sequences and, 1975, 1976 power control in (see also power control), 461-462, 1982-88 principles of, 2874-76 pseudonoise sequences and, 459 radio resource management and, 2090, 2091-93 satellite communications and, 879, 881, 1231-32, 1231 scrambling codes in, 2876, 2877 serially concatenated coding and, 2176-77, 2176 shallow water acoustic networks and, 2208, 2209, 2215 signal to noise and interference ratio and, 458, 459-460 signature sequences in, 2274-85, 2275 software radio and, 2312-13, 2312, 2314, 2316 speech coding/synthesis and, 2354 spread spectrum and, 458, 2276-78, 2400 spreading factor in, 459 spreading in, orthogonal and nonorthogonal, 2874-75 synchronization and, 2479-81, 2479 synchronous CDMA and, 1096 TD/CDMA and, 2589-90, 2592 ternary sequences and, 2536-47 time division multiple access and, 2586, 2590, 2907-09 turbo product coding and, 2727-37 two layer spreading in, 2875-76, 2875 ultrawideband radio and, 2754-62 universal mobile telecommunications system and, 2873-74 UTRAN and, 2873-74 Walsh-Hadamard sequences in, 2282-83

3012 INDEX

code division multiple access (see also cdma2000) (continued) wideband CDMA, 733-734, 1096, 1104-05, 1986, 2116, 2282-83, 2400, 2873-83, 2950-51, 2954 wireless local loop and, 2950-51, 2955 wireless multiuser communications systems and, 1602, 1608, 1609, 1615 Code Division Testbed project, 397 code excited linear prediction, 2820-29, 2824-28 codevector-based approach to training in quantization and, 2129 codeword, codeword polynomial, BCH coding, binary, and, 244 codewords constrained coding techniques for data storage and, 573 cyclic coding and, 618 sequential decoding of convolutional coding and, 2141 coding, synchronization and, 2480-81 coding (Reed-Solomon) for magnetic recording channels bit error rate in, 473 block error rate in, 473 block missynchronization detection in, 471-472 cyclic coding in, 469 error correcting coding in, 466-467, 466, 470, 472-474 error detecting coding in, separate vs. embedded, 474 error rate definitions for, 473 hard decision decoding algorithms for, 475 interleaving vs. noninterleaving in, 472 large sector size and, 475 linear coding in, 469 performance and 472-474 redundant array of independent disks and, 474-475 Reed-Solomon coding and, 467-475 soft bit error rate in, 474 soft decision decoding algorithms for, 475 symbol error rate in, 473 systematic coding in, 469 tape drive ECC and, 474 coding distance profile, 2160 coding excited linear prediction, 41, 1266-67, 1302-05, 1303, 2348–49, 2349, 2372, 2382 coding for magnetic recording channels, 466-476, 466 coding polynomial, cyclic coding and, 618 coding puncture rates, Universal Mobile Telecommunications System and, 386 coding tracking, in synchronization and, 2480 coding tree, sequential decoding of convolutional coding and, 2142, 2143, 2149 coding, graphs, cycles in, 1315 cognitive radio, 2307 coherence bandwidth, space-time coding and, 2327 coherent detection minimum shift keying and, 1468-70, 1469, 1470 optical fiber systems and, 1848 optical transceivers and, 1834-35, 1834 orthogonal frequency division multiplexing and, 1877 - 78pulse amplitude modulation and, 2026 wireless infrared communications and, 2926 coherent processing, in acoustic modems for underwater communications, 16-17 coherent receivers, optical communications systems and, 1484, 1486-88, 1486, 1487, 1488 collimation, in parabolic and reflector antennas and, 2082 collision domains, Ethernet and, 1281 collision warning systems, intelligent transportation systems and, 503 collision (see also media access control), 315, 547, 1280 1347 ALOHA protocol and, 130-131 carrier sense multiple access and, 343-346, 343 media access control and, 1342-1349 traffic engineering and, 500 colocation method, antenna modeling and, 174 color space, image and video coding and, 1026 colored background noise, powerline communications and, 2001

column distance, convolutional coding and, 603, 2142, 2158 comb filtering, automatic speech recognition and, 2378 comb line, microstrip/microstrip patch antennas and arrays in, 1374, 1376 combicoding, optical recording and, 581 combinatorial design, low density parity check coding and, 1316 common gateway interface, 2900 common object request broker architecture, 726-728, 2304, 2310 common open policy service, 1656 common packet channel switching, admission control and, 123 communication protocols, 538-556 communication satellite onboard processing (see also satellite communications), 476-485 access control in, 482 adaptive antennas and, 479-480 add drop multiplexers and, 482 amplifiers in, 477 antenna beam switching in, 478-479 antennas for, 477 automatic gain control in, 477 bandwidth limited case in, 478 beamforming in, 480 control link use in, 482 conventional (nonprocessing) satellite and, 476-477, 476 demodulation-remodulation in, 480-482, 480, 481 examples of systems using, 483-484 frequencies in, uplink vs downlink, 476 frequency reuse in, 479 geosynchronous satellite, total link capacity of, 478-479, 479 ground processing tradeoffs of, 483, 483 interconnected spot beams in, 479, 480 interference and, 477-478, 478 limitations of conventional satellite architectures and, 477-478 multiple access systems and, 477, 477 multiplexing and switching in, 482 packet switching and, 482-483 power limited case in, 478 security and survivability in, 482 spread spectrum and despreading, 482 system response time in, 482 translating repeater and, 476, 476 transponders and, 476 uplink interference and, 477-478, 478 user interconnection and, 478 communication security, 1651 communication system traffic engineering (see traffic engineering) communications for intelligent transportation systems (see intelligent transportation systems) community access TV, 512-527, 2653 AM systems in, 518-519, 519 amplifiers for, 512, 517 asynchronous transfer mode and digital video in, 524 attenuation-frequency response in coax and, 517-518, 517 beat noise/distortion in, 514 binary convolutional coder in, 526-527, 526 broadband and, 2668-71 carrier to noise ratio in, 515-517, 520 channel allocation in, 513-514 coaxial cable for, 517-618 composite triple beat in, 514 compressed video in, 522, 525 cross modulation in, 516-517 data communication using, 512, 524 data over cable service interface specification in, 524, 524 dBmV values in, application of, 514 digital transmission in, 522 digital video standards for, 524-527 evolution and history of, 513-514, 513 FM systems, 519-522 forward error correction and digital video in, 524, 525-527, **525**

frequency division multiple access, 523 gain in, 517 guard bands in, 523 headend in, 512, 513 hybrid fiber coax systems in, 512, 518-522, 518 impedance matching in, 524 ingress noise in, 524 interleaving in, 526 intermodulation in, 512, 514 IP telephony and, 1177 last mile communications and, 512 layouts for, 513-514, 513 link budget for AM systems in, 518-519, 522 microwave signals in, 513 MPEG-2 transport framing of digital video in, 525 noise in, 512, 514-517, 515, 523-524 NTSC standards and, 522 oversampling in, 522 picture ratings for, 515 powerline communications and vs., 1998 pseudorandom noise (PN) in, 526 quadrature amplitude modulation and digital video in, 524, 525, 526-527, 526 randomization in, 526 Reed-Solomon coding in, 526 sampling in, 522 satellite systems and, 514 signal to noise ratio in, 514, 515-522, 526 splitters in, 519, 520 spread spectrum and, 2399 supertrunks in, 512 taps in, 518, 518 thermal noise in, 514-515, 523-524 time division multiple access and, 523 trellis coding in, 526 two-way systems in, 522-524, 523 video transmission subsystem in, 520, 520 voice communication using, 512, 523-524 compact disc (see also CDROM; optical memories), 579-581, 1319, 1735-36, 1736, 1735 constrained coding techniques for data storage and, 579-581 CIRC encoders for, 627-628, 627 cyclic coding and, 626-628 Reed-Solomon coding and, 626 compact HTML, mobility portals and, 2193 companders, 527-530, 528 A law, 529, 529 pulse coding modulation and, 527-530, 528speech and, 529 u law, 529, **529** waveform coding and, 2834 compatibility issues, in cdma2000 and, 367 compensation of nonlinear distortion in RF power amplifiers (see RF power amplifiers, nonlinear distortion in) competitive learning, neural networks and, 1678 complementary cumulative distribution functioin, 1945-46, 1946 complementary key coding, 2943 complementary sequences, Golay, 892-900 complementary slackness, in flow control and, 1629 complemented cycling registers in, 799 complemented summing registers in, 799 complexity barrier, in vector quantization and, 2126 composite baseband channel, in channel modeling, estimation, tracking, 399 composite capabilities/preference profiles, 2194 composite triple beat, 514 compression, 631-650, 632 arithmetic coding in, 636-638 bandwidth and, 631 BISDN and, 263 bit allocation in, 646 channel coding in, 631 community antenna TV and, 522, 525 companders and, 527-530 compression rate in, 633 differential coding and, 648 distortion bounds in, 640-641 distortion upper bound in, 641
compression (continued) distortion-rate function in, 640-641 entropy bounds in, 633-634 enumerative coding in, 635-636 fractal images and, 648 Gaussian memoryless sources in, 641 Hamming distortion in, 640 Huffman coding and, 634–635, 637, 1017–24 image and video coding and, 1028-29, 1030 image, 1062-73, 1075-76 iterative coding in, 635-636, 635 JPEG compression and, 1211-18 Karhunen-Loeve transform in, 648 Kraft's inequality in, 633, 635 lattice vector quantization in, 644 LBG algorithm in, 643-644 Lempel-Ziv coding in, 638-639 Lloyd-Max quantizers in, 642 lossless, 632-633, 2123, 2124 lossy, 632-633, 639-648, 2123, 2124 Marcelling-Fischer coding in, 646 Markov source in, 632, 634 mathematical description of source in, 632 memoryless source in, 632, 633-634, 641 modems and, 1496 multimedia over digital subscriber line and, 1570-71 nearest neighbor quantization in, 642 pointer encoding and, 638-639 prefix conditions in, 633 quantization and, 639 reliability and, 631 sampling and, 631-632 scalar quantization and, 641-642 Shannon-Fano coding in, 634-635 source coding in, 631 speech coding/synthesis and, 648 squared error distortion in, 640 subband coding and, 648 transform coding and, 646-648, 645 tree structured vector quantization in, 644 trellis coding and, 644-646, 645 in underwater acoustic communications, 36, 37 vector quantization and, 642-644 Viterbi algorithm and, 644 wireless IP telephony and, 2935-36, 2936 wireless packet data and, 2987 compression filters, chirp modulation and, 442 compression gain, chirp modulation and, 443 compression of data, 371 compression rate, 633 computational science, 1675 computer communication protocols, 538-556 Comsat, 268, 876 concatenated convolutional coding and iterative decoding (see also convolutional coding) 556-570 Bahl-Cocke-Jelinek-Raviv decoding and, 556, 561-564. 564 bit error rate and, 559-560, 560 encoder structures for, 556-558, 556 flooring effect in, 560 interleavers in, 557-558 maximum likelihood decoding in, 558-560 parallel, 556, 557-560, 559, 564-567 puncturer in, 558 recursive systematic convolutional coding and, 556-557, 557 serial, 556, 557-560, 559, 567-569, 567 spectral thinning and, 558 turbo coding as, 556 concatenated multidimensional parity check coding, 1548–49, **1549** concave diffraction gratings, 1755 concave gratings, two-dimensional, 1754 concentric ring circular antenna array, 150-151 condenser microphone, transducers (acoustic) and, 34, 34 conditional joint probability, maximum likelihood estimation and, 1338 conditional mean estimator, 1340-1341, 1340 conditional statistical optimization, packet rate adaptive

mobile receivers and, 1892 cone penetrometer using underwater acoustic modem, 19-20, 20 conferencing, session initiation protocol (SIP) and, 2202 confidentiality of data, 1151-52, 1648 confinement factor, lasers and, 1777, 1778 conformal antenna arrays and, 142, 152-153, 152, 169 confusion concept, cryptography and, 606 congestion avoidance and control (see also flow control; traffic engineering), 551-552, 1661-63 additive increase multiplicative decrease in, 1662 admission control and, 112 explicit congestion notification in, 1662-63, 1662 explicit rate feedback in, 1663 explicit rate indication for congestion avoidance in, 1663 flow control and, 1625-31, 1653 forward acknowledgement in, 1662 multimedia networks and, 1566 multiprotocol label switching and, 1594, 1599 packet switched networks and, 1907, 1910 preventive, 112 reactive, 112 real time control protocol in, 1662 satellite communications and, 2120 selective acknowledgement in, 1662 shallow water acoustic networks and, 2211 streaming video and, 2438-39 TCP friendly rate control in, 1662 traffic engineering and, 491, 491 transmission control protocol (TCP) and, 553-554, 1661-62, 2610-11, 2611 transport protocols for optical networks and, 2616-17 user datagram protocol and, 1662 conical conformal antenna arrays and, 152-153 conjugacy classes, cyclic coding and, 618 conjugate gradient method, adaptive antenna arrays and, 72-73 conjugate structure CELP, 1304, 1306 connection admission control, 205, 1625, 2004 connection control or control plane, ATM and, 200 connection oriented networks ATM and, 265, 550 Bluetooth and, 313 packet switched networks and, 1909-10, 1909 connection polynomial, in BCH (nonbinary) and Reed-Solomon coding, 257-258 connectionless networks Bluetooth and, 313 IP networks and, 269 packet switched networks and, 1909-10, 1909 connections, in flow control, traffic management and, 1653 connectivity, media access control and, 5, 1343 connectors, optical fiber and, 1707 constant angular velocity, CDROM and, 1735 constant bit rate, 123, 206, 266, 551, 552-553, 1658, 1663 constant linear velocity drives, CDROM and, 1735 constant modulus algorithms, 292, 1614 constant-modulus algorithm, equalizers and, 92 constellation labeling, bit interleaved coded modulation and, 279 constellation shaping, shell mapping and, 2221-22, 2221 constituent coding, serially concatenated coding and, 2164 constrained coding for data storage, 570-584 ACH algorithm in, 578, 579, 581 approximate eigenvectors in, 576-577 biphase coding in, 576 block coding in, 576-579 Blu-Ray Disc in, 579 bounded delay encodable coding in, 578 channel encoding and, 570 characteristic equation in, 574 coding construction methods in, 576-579 coding rate and capacity in, 573-575 combicoding in, 581 constrained sequences or codewords in, 573

DC control in, 579-581, 580

detection window or timing window in, 573 deterministic systems in, 575 efficiency in, 573 eight to fourteen modulation in, 579 encoder/decoder in, 573, 575-576, 575 enumerative methods in, 577 error correcting coding in, 570, 579 error propagation in, 576 finite local coanticipation in, 575 finite state transition diagram in, 573 finite type, 571 fixed length prinicipal state coding in, 577 follower sets in, 575 frames, frame headers in, 576 frequency domain constraints in, 579 frequency modulation in, 576 global and interleaved constraints in, 582 guided scrambling in, 579 jitter or mark edge noise in, 573 Kraft-McMillan inequality in, 577-578 lookahead and history in, 578 maximum transition run constraints in, 581-582 memory and anticipation in, 575 merging bits, merging rules in, 576 Miller coding in, 576 modified frequency modulation coding in, 576 modulation coding in, 570, 573 modulation transfer function in, 572-573, 572 multitrack coding in, 582 non return to zero in, 570-571, 581 non return to zero inverse in, 570-571, 579, 580 optical recording and, 579-581 parity check coding in, 581 partial response maximum likelihood in, 582 Perron-Frobenius theory and, 574 phase locked loop in, 571 pits and lands in, 570 power spectral density function in, 579 principal state method in, 577–578 Reed–Solomon coding and, 576 run length limited in, 571-573, 571, 579-581 running digital sum in, 579 Shannon cover in, 575 Shannon's law and, 573 sliding block decoders in, 575 sofic systems in, 573-575 source and channel bits in, 573, 575 spectral null constraints in, 579 state combination in, 578 state merging in, 575 state splitting, weighted vs. consistent, 578-579, 579 subset construction in, 575 substitution coding in, 581 substitution method in, 578 subword closed systems in, 573 synchronization in, 576 time domain constraints in, 579 time varying coding in, 581 two dimensional constraints in, 582 variable length principal state coding in, 578 window size and, 575 constrained sequences, constrained coding techniques for data storage and, 573 constrained source coding, transform coding and, 2594–96, **2595** constrained tree ad hoc wireless networks, 2891 constraint-based label distribution protocol, 1596 constraint-based routing flow control, traffic management and, 1654-55 multimedia networks and, 1568 constraint length, sequential decoding of convolutional coding and, 2142 containers, synchronous digital hierarchy and, 2498 content protection for recordable media, 1738 contention, ATM and, 201 contention-based protocols, media access control and, 1343, 1346-1347 continuity checking, ATM and, 207 continuous phase frequency modulation, 1457 continuous phase frequency shift keying, 593-598 continuous phase modulation and, 593-598

continuous phase frequency shift keying (continued) definition of, 593-594 error detection and correction in, 594-598, 596 filters in, 596 frequency modulation and, 593-594 frequency shift keying and, 593 minimum shift keying and, 593-598, 1457 noncoherent structures and performance in, 596-598, 598 phase shift keying and, 593 phase trellis in, 597 power spectra of digitally modulated signals and, 1989, 1991 receivers for, and error rate performance of, 594-598, 595 signal to noise ratio in, 596-597, 597 trajectories of, 593, 593 transmitted spectral properties of, 594, 594 Viterbi algorithm and, 597 continuous phase modulation, 584-593, 710, 718-719 a posteriori probability algorithm in, 2180-81 additive white Gaussian noise in, 589 bandwidth and, 589-590, 2180 bit error rates and, 2180, 2181-89, 2183 continuous phase frequency shift keying and, 593-598 convolution coding and, 2180 convolutional coding and, 591 correlative states in, 586 detection and error probability in, 587-590 duobinary frequency shift keying and, 585 error detection and correction in, 2182-84, 2183 fast frequency shift keying and, 584 filtered, 592 free distance in, 589 full- and partial-response techniques in, 585-586, 586-587 Gaussian minimum shift keying and, 584-593 generalized tamed frequency modulation and, 585 generation of signals and transmitters in, 590, 590 Hamming distance and, 2182 iterative decoding in, 2180-81 minimum shift keying and, 584-593, 1457-59, 1458, 2182 partially coherent detection and, 591 phase offset in, 588 power spectra for, 587, 588, 589 power spectra of digitally modulated signals and, 1989, 1991-94, 1992, 1994, 1994 quadature phase shift keying and, 589 raised cosine modulation and, 585 receivers for, 591, 591, 592 recursive systematic convolutional coding and, 2182 satellite communications and, 1225 serially concatenated coding and, 2173-75, 2174, 2175, 2179-90, 2180 Shannon's theory and, 592 signal classes of, 585-587 signal to noise ratio and, 588-589, 2181-83, 2189 spectrally raised cosine modulation and, 585 synchronization and, 2473-85 tamed frequency modulation and, 584-593 trellis coding and, 587, 590 Viterbi detectors and, 591 continuous shift keying, 1473-74 continuous speech recognition, 2377 continuous time, in channel modeling, estimation, tracking, 398-399, 399 continuous wave lasers, free space optics and, 1852-53 continuously varying slope delta modulation, 2343, 2356 control channel, 1478, 1803 control links, satellite onboard processing and, 482 control plane, 264, 1798 control signals, in underwater acoustic communications, 37 controlled cell transfer, ATM and, 267 conventional (matched filter) receiver, 97-98, 106

conventional double sideband AM, 133, 134-135, 135 convergence

power control and, 1985

rate distortion theory and, 2075 serially concatenated coding and, nonconvergence region, 2166-67, 2167 convergence zone, in underwater acoustic communications, 38 conversion, sigma delta, 2227-47, 2228 convolutional coding (see also concatenated convolutional coding and iterative decoding), 556-570, 598-606, 2040-64, 2179 additive white Gaussian noise and, 599, 601-602, 605 applications for, 598 bit error rate in, 599, 602-605 block error rate in, 602-605 bounds on bit error rate in, 604-605 catastrophic encoders for, 603-604, 604 catastrophic encoders in, 598 channel coding and, 598 circular recursive systematic convolutional coding and, 2709, 2709 column distance in, 603 continuous phase modulation and, 591, 2180 decision depth in, 603, 605 decoding of, 600-602, 600, 2040-64, 2140 encoder structure in, 599-600, 599 equivalent encoders for, 599-600 error correction coding and, 598 feedback and feedforward encoders for, 599-600, 599 finite traceback Viterbi decoding in, 602, 602 free distance in, 598, 602-604, 603 generating function in, 604-605 Hamming distance in, 602-604, 603 hard vs. soft decoding in, 601-602, 602 interleaving and, 1141-51, 1142-1149 IS95 cellular telephone standard and, 353 maximum a posteriori decoders and, 600 maximum likelihood decoders and, 600 minimal encoders for, 600 packet binary convolutional coding and, 2946 parallel concatenated convolutional coding, 2710 path metrics of Viterbi algorithm in, 600 punctured, high rate, 979-993 recursive systematic convolution coding and, 2705-07, 2706 satellite communications and, 1229-30, 1230 sequential decoders and, 600, 2040-64 serially concatenated coding for CPM and, 2185-86, 2185 Shannon's theory and, 605 shift registers and, 598 signal to noise ratio in, 602-604, 605 speech coding/synthesis and, 2355 split state diagrams for, 604-605, 604 tail biting, 2511-16, 2513 terminating, 2511-13 threshold coding and, 2581-83, 2582 trellis coding and, 2645-46, 2645 trellis diagrams in, 600, 600, 601 turbo coding and, 600 union bounds in, 598-599, 605 Viterbi algorithm and, 598-602, 601, 2816-17, 2816 wireless multiuser communications systems and, 1609–10, **1609**, **1610** convolvers, surface acoustic wave filters and, 2455-56, 2455, 2456, 2460 cooling schedule, quantization and, 2130 Cooperation of Field of Scientific and Technical Research projects, 397 coplanar microstrip feed line, 1362-1363, 1362 coplanar waveguide, active antennas and, 51-52, 52, 64-66, 1354 copper media, 1706 free space optics and, 1851 local area networks and, 1283 orthogonal frequency division multiplexing and, 1867 COPS protocol, admission control and, 116 cordless telephone antennas, 189

Bluetooth and, 307

indoor propagation models for, 2012-21

1715 core assisted mesh protocol, 2891 core-based tree, 1535 core extraction distributed ad hoc routing, 2890 core networks, optical fiber systems and, 1840, 1841 core routers burst switching networks and, 1802, 1804, 1804 differentiated services and, 669 corner illuminated cells, 450, 450 corner reflector, 191, 191 corrective filters, 1723 correlated Gaussian sequences, random number generation and, 2292-93 correlation adaptive receivers for spread-spectrum system and, estimation in, 107-108 cellular communications channels and, 397 optical fiber systems and, 702-709, 703, 705, 708, 708-709 pulse amplitude modulation and, 2026 pulse position modulation and, 2035 cost functions, maximum likelihood estimation and, 1340 cost or objective function, adaptive receivers for spreadspectrum system and, 99-100 Costas loop, 2028, 2028, 2054, 2054 costing, price of links, flow control and, 1628-29 counterpropagating gratings (see also Bragg gratings), 1723, **1723** coupled mode theory, optical filters and, 1728-29, 1729 couplers and coupling active antennas and, 52 adaptive antenna arrays and, 68-69, 70, 73-77, 74, **75, 76, 77,** 74 antenna arrays and, 160, 164-166, 165, 166 antenna modeling and, integrals in, 174 high frequency communications and, 951 microstrip/microstrip patch antennas and, 1370-1371, 1371, 1372, 1373 optical fiber and, 1697-1700, 1697-1700, 1707, 1715, 1715 powerline communications and, 1998-99 coupling coefficient, optical couplers and, 1698 coupling loss, diffraction gratings and, 1751-52, 1753 covariance, adaptive antenna arrays and, 68 coverage area or footprint cell planning in wireless networks and, 372 satellite communications and, 1249, 2111 wireless LANs and, 2678 wireless packet data and, 2985 wireless sensor networks and, 2995 Cramer-Rao bound, 402, 403, 405, 405, 1339 Cramer-Rao lower bound, 2055 credit-based control schemes, ATM and, 551 credit weighted algorithms, medium access control and, 1556 creeper algorithm, sequential decoding of convolutional coding and, 2154 critical frequency, 2065, 2066 cross correlation diversity and, 729 feedback shift registers and, 796 Gold sequences and, 901-902 polyphase sequences and, 1975 pulse position modulation and, 2042 signature sequence for CDMA and, 2276-85 ternary sequences and, 2542-43 cross coupled IQ transmitter, minimum shift keying and, 1467, **1467** cross gain modulation signal quality monitoring and, 2273 wavelength division multiplexing and, 756 cross modulation, community antenna TV and, 516-517 cross phase modulation optical communications systems and, 1490-91, 1490, 1684, 1686–87, **1687**, 1712, 1844, 1846 solitons and, intrachannel, 1769-70 wavelength division multiplexing and, 756 cross polarization, local multipoint distribution services and, 1277, 1277

core, optical fiber and, 434, 435, 1708, 1708, 1714,

IS95 cellular telephone standard and, 353

cross validated minimum output variance rule, 1895, 1898 crossbar switch, ATM and, 202-203, 202 crossed dipole antenna, 199 crossed drooping dipole antenna, 197, 197 crossed slot antenna, 199 crossover probability, automatic repeat request and, 230, 230 cross-spectral reduced-rank method, adaptive receivers for spread-spectrum system and, 104 crosstalk cellular communications channels and, 397 diffraction gratings and, 1752 digital magnetic recording channel and, M7, 1325 optical couplers and, 1699 optical cross connects/switches and, 1784-85 optical fiber systems and, 1843 optical signal regeneration and, 1759 photonic analog to digital conversion and, 1965 very high speed DSL and, 2786, 2798-2800, 2803-05, 2805 cryptoanalysis, 607 cryptography, 606-616, 1151-52 advanced encryption standard in, 606, 608, 610, 1152, 1648 asymmetric key/public key, 606, 607, 611-612, 1152 authentication and, 613-614 authentication in, 606, 607, 611 block ciphers in, 607-609, 607 Blum Blum Shub random number generator in, 615 brute force attacks on, 607-608 cable modems and, 335 cipher feedback in, 607 complexity of, 609-610 confusion concept in, 606 cryptoanalysis and, 607 data encryption standard in, 606, 607-608, 607, 1152, 1648 Diffie Hellman coding and, 606, 610-611, 612, 614, 1152 diffusion concept in, 606 digital signature algorithm and, 612-613 digital signatures in, 606, 607, 612-613, 1649 discrete logarithm problem in, 609-610 double random phase encryption in, 2132-33 El Gamal encryption in, 612, 613, 1649 electronic cash and, 615 electronic codebook in, 607 elliptic curves in, 610, 610 elliptical curves and, 613 factor bases in, 610 Federal Information Processing Standards and, 606 Fiat Shamir identification protocol in, 614 general packet radio service and, 875 H.324 standard and, 922 hash functions in, 612-613, 1152 Hasse-Weil theorem in, 610 holographic memory/optical storage and, 2132 index calculus method in, 610 key distribution center, 1152 key exchange in, 610-611 keyed hash MAC in, 613 man in the middle/person in the middle and, 611 manipulation detection coding in, 613 message authentication coding in, 613 Miller Rabin method in, 614-615 one way functions in, 606, 609-610, 1152 output feedback in, 607 Pocklington's theorem in, 615 point doubling in, 610 prime number generation and, 607, 614-615 private key encryption, 1152 public key encryption, 1152 public key infrastructure in, 614 quantum computation vs., 615-616 Rabin encryption in, 611-612 random number generation and, 607, 614-615 RSA algorithm in, 606, 611, 615, 1152, 1649 Schnorr identification protocol and, 614 secure hash algorithm and, 612 security and, 1647, 1648-49, 1651

standards and documentation for (NIST), 606 state matrices in, 608 stream ciphers in, 607-609, 608 symmetric key/private key, 606, 607-609, 607, 1152 trap door one way functions in, 606, 609-610 XORing in, 608-609 crystal radio sets, 1477 CSMA/IS95 cellular telephone standard (see IS95 cellular telephone standard), 347 cumulant matching, blind equalizers and, 293-294 cumulative density function, 451, 1946 current density, loop antennas and, 1294 customer edge, 1599 customized application for mobile network, 908, 1101-08 cutback method testing, optical fiber and, 436 cutoff frequency, 1395, 1396, 2110 cutoff rates, sequential decoding of convolutional coding and, 2158 cycle covers, 1638-39, 1638 cycles, low density parity check coding and, 1315 cyclic block coding, BCH coding, binary, and, 243-244 cyclic coding (see also BCH coding; Golay coding; Reed-Solomon coding), 616-630 applications for, 617, 626-629 basic properties of, 618-619 BCH coding and, 616, 621-626 BCH coding, binary, and, 243-252 Berlekamp decoding algorithm, 624-625 CCSDS standards and, 629 Chien search and, 617 CIRC encoders for, 627-628, 627 codewords in, 618 coding polynomial in, 618 compact-disk players using, 626-628 conjugacy classes in, 618 cyclotomic cosets in, 618 decoding in, 617 deep space telecommunications and, 628-629 direct solution algorithm and, 617 elementary symmetric functions in, 623 error correcting coding and, 617 error detection and correction in, 620 error locators and, 623 error trapping decoder for, 617 Euclid's algorithm and, 617 Galois fields in, 617, 618, 620 general theory of, 617-620 generating functions in, 623 generator polynomials in, 618 Golay coding and, 616, 620-621, 628-629, 885-892 Hamming coding and, 617 history and development of, 616-617 linear feedback shift register in, 619-620, 625-626, 625, 626 Massey-Berlekamp decoding algorithm, 617, 625-626 maximum likelihood algorithm and, 620 minimal polynomials in, 617-618 Newton's identities and, 623, 625 power sum symmetric functions in, 623 product coding as, 1539-40 quadratic residue coding and, 616-617, 620-621 Reed-Muller system in, 628 Reed-Solomon coding and, 469, 616, 622-626, 629 Shannon's theory and, 617 shift register encoders/decoders for, 619-620 syndrome decoder for, 620 syndrome equations in, 623 systematic encoding in, 619, 619 cyclic Hadamard difference sets, feedback shift registers and, 790, 795 cyclic redundancy check ATM and, 264 automatic repeat request and, 225-226 BCH coding, binary, and, 245 Bluetooth and, 312 cdma2000 and, 367 Ethernet and, 1503 failure and fault detection/recovery in, 1633-34

smoothing in, 610

modems and, 1495 shallow water acoustic networks and, 2207 speech coding/synthesis and, 2355 cyclic reservation multiple access, 1558 cyclic suffix, very high speed DSL and, 2795-96, 2796, 2797 cyclostationary process power spectra of digitally modulated signals and, 1989-90 pulse amplitude modulation and, 2023-24 pulse position modulation and, 2035 cyclotomic cosets, 622, 618 cylindrical antenna array, 151-152, 152 cylindrical conformal antenna arrays and, 152, 152 Czerny–Turner spectrograph, diffraction gratings and, 1751, **1752**, 1754 D/T (see propagation time) damping, active antenna, 60 dark current noise, in optical fiber systems, 1843 Darlington filters, 686 data burst channels, burst switching networks, 1803 data communication, community antenna TV and, 512, 524 H.324 standard, 918-929, 919 data communication equipment, modems, 1495, 1496 data compression (see compression) data confidentiality, 1648 data directed channel tracking, 415-418 data encryption standard, 335, 606, 607-608, 607, 1152, 1648 data frames, peak to average power ratio, 1945 data fusion, location in wireless systems, 2961-64 data hiding, image processing, 1077 data integrity, 1648, 1649 data link control layer, 1281 OSI reference model, 539-540, 544 packet switched networks and, 1910-11 shallow water acoustic networks and, 2207 data link control protocols, 544-547, 952 data networks, admission control, 116-117 data origin authentication, 1647 data over cable service interface specification, 272 broadband wireless access and, 317, 2670-2671, 2670 cable modems and, 324, 327, 333, 334 community antenna TV and, 524, 524 IP telephony and, 1177 modems and, 515 data preprocessors, sigma delta converters, 2243, 2244 data rates adaptive receivers for spread-spectrum system and, multiple, 108-109, 109 broadband and, 2654-55 optical fiber and, 1714 powerline communications and, 1995 space-time coding and, 2324 underw/in underwater acoustic communications, 37 wireless packet data and, 2982 data search information, digital versatile disc, 1738 data storage (see also constrained coding techniques for; magnetic storage systems), 570-584, 1319 data terminal equipment, modems, 1495, 1496 data transfer rates, hard disk drives, 1322 data/information rates, pulse position modulation, 2039-2040, 2040 datagrams IP networks and, 269, 542, 543 shallow water acoustic networks and, 2211 Datakit virtual circuit, 264 datalogging, in acoustic modems for underwater communications, 18 Datasonics, 24 DATS telemetry system, 24 day to day variation, traffic engineering, 488 daytime measurement, radiowave propagation, 2064 DC canceller, sigma delta converters, 2243-45, 2244, 2245, 2246 DC control, optical recording, 579-581, 580 DC drift, optical modulators, 1746-47, 1746

DCS1800, 370 De Bruijin sequences, feedback shift registers (FSR), 790, 795 dead zones, 215 decametric (see high frequency) decision depth, 603, 605, 2648 decision directed algorithms, blind equalizers, 291 decision directed feedback equalizers, tapped delay line equalizers, 1693 decision feedback equalizers, 16, 81, 87-89, 88, 89, 286 acoustic telemetry in, 26 adaptive receivers for spread-spectrum system and, 105, 105 blind equalizers and, 289, 290, 292 cable modems and, 330 in channel modeling, estimation, tracking, 416, 417-418, 417 magnetic recording systems and, 2262-63 polarization mode dispersion and vs., 1973-74, 1973 space-time coding and, 2328 tapped delay line equalizers and, 1688-89, 1692 tropospheric scatter communications and, 2701-02, 2701, 2702 very high speed DSL and, 2802, 2803 decision regions, permutation coding, 1955-56 decision-directed algorithms, adaptive receivers for spread-spectrum system, 105, 105 decision-directed mode, equalizers, 82 decoders, decoding in BCH (nonbinary) and Reed-Solomon coding, 469-470, 622-626 constrained coding techniques for data storage and, 573, 575-576, 575 convolutional coding and, 600-602, 600 cyclic coding and, 617, 619-620 linear predictive coding and, 1264 permutation coding and, 1955 Reed-Solomon coding, 622-626 soft output algorithms for, 2295-2304 threshold type, 2579-85 trellis coded modulation and, 2627 turbo coding and, 2705, 2705, 2713-14, 2713 turbo trellis coded modulation and, 2738, 2743-47, 2744, 2745, 2746 ultrawideband radio and, 2755-58 vector quantization and, 2125 Viterbi algorithm and, 2815-19, 2815 wireless multiuser communications systems and, 1618-19 decorrelating detector, 98, 463-465, 463, 1616 decorrelation, image compression, 1063-65 decorrelation filters, acoustic echo cancellation, 7, 8, 7 decryption system for holographic memory/optical storage, 2137-38, 2137 dedicated short range communications, 506 deemphasis filtering, 821-823 deep reactive ion etching, microelectromechanical systems, 4, 1352 deep space telecommunications, cyclic coding, 628-629 Defense Communications Agency, 268 Defense Satellite Communication System, 483 deficit round robin, multimedia networks, 1565 degeneracy factor, optical fiber systems, 1846 degrees of freedom, adaptive antenna arrays, 73 delay broadband and, 2655 fading and, 783 flow control, traffic management and, 1627, 1653, 1660 indoor propagation models and, 2017 IP telephony and, 1172-82, 1173 multimedia networks and, 1562 optical signal regeneration and, 1761, 1761 path loss and, 1937 power control and, 1986 satellite communications and, 879, 1250-51, 2112 in underwater acoustic communications, 38-40, 39 wireless packet data and, 2984 delay coefficients, acoustic echo cancellation, 10 delay lines, surface acoustic wave filters, 2448-49, 2450 - 52delay power spectrum of channel, wireless, 1604

delay spread, 394, 395, 410 delayed decision feedback sequence estimation, 81 delayed interference devices, optical signal regeneration, 1763 1763 delta modulation, 648, 2835 delta rule, in neural networks, 1677-78 demand assignment multiple access, 879, 956 demilitarized zones, 1650 demodulation/demodulators additive white Gaussian noise and, 7, 1335 amplitude modulation and, 137-140 chann/in channel modeling, estimation, tracking, narrowband, 411 digital phase modulation and, 709-719 double sideband suppressed carrier AM, 140, 140 filters in, 7, 1335 matched filters in, 1335-1338, 1336 maximum a posteriori algorithm in, 1335 pulse amplitude modulation and, 2024-30 pulse position modulation and, 2036-39 quadrature amplitude modulation and, 2047 single sideband AM, 140, 140 vestigial sideband AM, 140 wireless infrared communications and, 2927-27 demultiplexers/demultiplexing, 540 diffraction gratings and, 1751-52 optical communications systems and, 1484, 1748-59, 1748 wireless transceivers, multi-antenna and, 1580 denial/degradation of service, 1646, 1647 dense wavelength division multiplexing, 748-757, 749, 2461-72, 2862 BISDN and, 273 Gigabit Ethernet and, 1509 optical cross connects/switches and, 1701, 1783, 1797 optical fiber and, 1709, 1720-21, 1797 signal quality monitoring and, 2271-72, 2273 solitons and, 1771 turbo coding and, 2728-37 density evolution, low density parity check coding, 1315 density function, maximum likelihood estimation, 1338 density in SPC coding, 1540 density of media, sound propagation, 30-31, 31 depolarization, millimeter wave propagation, 1272, 1439-40, 1445 descent algorithm in quantization, 2129 design distance, BCH coding, binary, 245 destination allocation protocol, 1552 destination sequence distance vector, 2211, 2888 detection/detectors in channel modeling, estimation, tracking, 398-408 chirp modulation and, postdetection integrator for, 445 magnetic recording systems and, 2258-63 pulse amplitude modulation and, 2024-30 quadrature amplitude modulation and, 2047 detection window, constrained coding techniques for data storage, 573 deterministic approach to admission control, 117 deterministic channel modeling and estimation, 405, 411-412 deterministic maximum likelihood algorithm, 405, 406 deterministic models, indoor propagation models, 2018-20 deterministic systems, constrained coding techniques for data storage, 575 DFH-3 feed waveguides, 1392, 1392 diagnostic acceptability measure, 2352 diagnostic alliteration test, 2352 diagnostic rhyme test, 2352 dichroic antenna arrays, 142 dielectric leaky wave antennas, 1244-45, 1244, 1245 dielectric filled waveguide, 1401-05, 1401, 1403, 1404, 1411-16, 1413-16 dielectric losses in waveguides, 1405, 1405 dielectric thin film stack interference filters, 1723, 1723, 1726-27, 1726, 1749 difference frequency generation laser, 1853 difference set cyclic coding, 802 differential approach to antenna modeling, 170

differential coding, compression, 648 differential delay, optical signal regeneration, 1761, 1761 differential detection, orthogonal frequency division multiplexing, 1877 differential group delay, 1492-93, 1970-71, 1971 differential least-squares algorithm, 107 differential mapping, image and video coding, 1030 differential PCM image and video coding and, 1037-38 speech coding/synthesis and, 2342-43, 2343 differential phase shift keying, 715–717, 716 acoustic telemetry in, 23 adaptive receivers for spread-spectrum system and, 107 modems and, 1497 optical transceivers and, 1825, 1830-31, 1831 satellite communications and, 1225, 1225 ultrawideband radio and, 2755 underw/in underwater acoustic communications, 41 differential pulse code modulation, 648, 2835, 2835 differential QPSK, 444, 717-718, 718, 1831-1832, 1832 differential shift keyed modulation, 422, 425-427, 426 differentially coherent detection, 1470-71, 1471, 1470 differentiated service, 668-77, 669 adaptive marking for aggregated TCP flows in, 672-673, 673 admission control and, 114, 115 architecture for, 668-70 assured forwarding in, 270-271, 669, 670-673, 670, 671 assured rate in, 675 behavior aggregate in, 270 boundary routers in, 668-669 bulk handling in, 675 core routers in, 669 differentiated services code point and, 668, 1657 egress routers in, 668 expedited forwarding in, 271, 669-670, 673-675, 1657–58 flow control, traffic management and, 1654, 1657-58, 1658, 1659, 1660 forwarding in, 669-673, 1657-58 hybrid IntServ-DiffServ in, 271 ingress routers in, 668 IP networks and, 270-271 IP telephony and, 1180 jitter in, 674, 674 mobility portals and, 2195 multimedia networks and, 1568-69 multiprotocol label switching and, 1594, 1597, 1598 packet scale rate guarantee in, 674 per domain behavior in, 675 per hop behavior in, 270, 1657, 1658 QBone and, 674-675 quality of service, 668 relative, 675-676 service level agreements and, 270, 668-77 traffic conditioning agreement in, 270 transmission control protocol and, 672-673, 673 video streaming and, 675 virtual wire service in, 674 differentiated services code point, 668, 1568, 1657 Diffie Hellman coding, 606, 610-612, 614, 1152, 1156 diffraction cellular communications channels and, 393 diffraction gratings and, 1750 geometric theory of, 2018 geometric theory of diffraction and, 216 indoor propagation models and, 2013, 2018 millimeter wave propagation and, 1438-39, 1445 optical multiplexing and demultiplexing and, 1749 parabolic and reflector antennas and, 1924 path loss and, 1940 radiowave propagation and, 213-215, 214, 215, 215-216 unified theory of diffraction and, 216, 1936, 1942, 2018 diffraction gratings (see also Bragg gratings; optical multiplexing/demultiplexing), 1723-26, 1723, 1726. 1749-56

diffraction gratings (see also Bragg gratings; optical multiplexing/demultiplexing) (continued) acoustooptical gratings in, 1755-56, 1756 arrayed waveguide grating in, 1752-54, 1753 Bragg condition in, 1756 classification of, 1750-51 concave, 1755 coupling loss and, 1751–52, **1753** crosstalk and, 1752 Czerny-Turner spectrograph configuration in, 1751, **1752**, 1754 demultiplexer peformance and, 1751-52 diffraction in, 1750 Ebert-Fastie spectrograph configuration in, 1751, **1752**, 1754–55 far field intensity in, 1749-50 focal curves in, 1751 free space gratings in, 1754-55 holographic concave gratings in, 1755, 1755 lasers and, 1779 operation of, 1749-50, 1751 optical multiplexing and demultiplexing and, 1749-56, 1758 Rayleigh criterion in, 1750 retroreflectors in, 1755 Rowland circle in, 1751, 1752 spectrograph overview of, 1751, 1752 STIMAX free space grating, 1754-55 two dimensional concave gratings in, 1754 Vernier effect and, 1780 diffusion, in photodetectors, 999-1000 diffusion concept, in cryptography, 606 digital advanced mobile phone system, 397 digital audio broadcasting, 321, 677-686 advantages of, 678 amplitude modulation and, 679-80 audio coding in, 681-682 channel coding in, 683-684, 683 digital signal processing and, 677-678 distortion and, 677 Eureka 147 DAB system in, 680-685 frequency bands for, 679-80 frequency modulation and, 679-680 global positioning system and, 685 history and development of, 678-680 integrated services broadcasting system in, 680 interference and, 677 masking in, 682, 682 modulation in, 684 MPEG compression and, 682-683 multipath and, 677, 678 multipath and, 685-685, 685 narrow band digital broadcasting in, 680 network for, 684-685 orthogonal frequency division multiplexing and, 678-679, 684 satellite digital audio radio service and, 680 single frequency networks in, 679, 679 transmitters for, 681, 681 digital audio tape, 1319 digital audio/video broadcasting, 2481, 2941 gital audio-visual council, 318, 320 digital beamforming antenna arrays, 142, 163-164 digital cross connect, 1634 digital elevation models, 374, 2561 digital enhanced cordless telecommunications, 1096-1108, 1289, 1350, 2952, 2955 Digital Equipment Corporation, local area networks, 1279 digital filters, 686-702 digital magnetic recording channel, 1322-1326, 1326, 1327 channel identification in, 1326 channel in, 1323-1324 distortion in, 1325-1326 dropouts in, 1326 finite impulse response equalizers and, 1324 head noise in, 1325 intersymbol interference and, 1325, 1326 intertrack interference (crosstalk) in, 1325 Lorentzian transition response in, 1324-1325, 1324, 1328-1329, 1329

media noise in, 1325 M-H curve in, 1322-1326, 1326 normalized linear density in, 1324 partial erasure in, 1326 preamplifier noise in, 1325 pulse amplituide modulation in, 1323 read process in, 1323-1324 thermal asperity in, 1326 transition shift in, 1325, 1325 write process in, 1323-1324 digital modular radio, 2306 digital phase modulation, 709-719 additive white Gaussian noise, 709 amplitude shift keying in, 709-719 bandwidth, 709, 710 binary phase shift keying,710-711 bit error rate in, 709 continuous phase modulation in, 710 frequency shift keying in, 709-719 Nyquist criterion, 709 phase shift keying in, 709-719 power efficiency in, 709 quadrature phase shift keying and, 710, 711 digital signal processing in acoustic modems for underwater communications, 17. 17 acoustic telemetry in, 23 blind equalizers and, 296 digital audio broadcasting and, 677-678 digital filters and, 686-687 discrete multitone and, 740-741 in underwater acoustic communications, 36 shallow water acoustic networks and, 2207 simulation and, 2285 software radio and, 2304, 2306, 2316 synchronization and, 2472-85 underw/in underwater acoustic communications, 43-44 wireless MPEG 4 videocommunications and, 2973 digital signature algorithm, 612-613 digital signature standard, 1649 digital signatures (see also authentication), 218, 606, 607, 612-613, 1649 digital simultaneous voice and data, 2340 digital subscriber line, 272, 2653, 2654, 2779-2807, 2781, 2915 access multiplexer for, 272 asymmetric, 1570, 1571-72, 1571 blind equalizers and, 287, 296 broadband and, 2655, 2666-73, 2670 broadband wireless access and, 317 compression and, 1570-71 integrated services digital networks and, 1570 layered coding in, 1570–71 modems and, 1499–1500 MPEG compression and, 1571 multimedia over, 1570-79 powerline communications and, 1995 quality of service and, 1571 satellite communications and, 2121 unequal error protection coding and, 2767 digital subscriber line access multiplexer, 272, 2784 digital terrain elevation data, microwave, 2561 digital terrain models, cell planning in wireless networks, 372 digital to analog converter in acoustic modems for underwater communications, 17 cable modems and, 333-334 digital filters and, 686-687 frequency synthesizers and, 833-835, 834 modems and, 1495 orthogonal frequency division multiplexing and, 1871 sigma delta converters and, 2227–47, **2228** software radio and, 2305, 2306, 2308, 2313 speech coding/synthesis and, 2370 wireless multiuser communications systems and, 1609 digital transmission, telephone, 262 digital TV, very high speed DSL, 2780 digital versatile disk (see also optical memories), 1319, 1733-41, 1734, 1737

capacity of, 1737 content protection for recordable media, 1738 data search information in, 1738 DVD-ROM media in, 1738 high definition TV and, 1738 MEPG compression and, 1738 multilayered memory in, 1737 NTSC standards and, 1738 PAL standards and, 1738 presentation control information in, 1738 read process in, 1737 recordable DVD-R media, 1738 standards for, 1737-38 write process in, 1738 digital video broadcasting, 2112, 2549, 2550 broadband wireless access and, 321 community antenna TV and, standards for, 524-527 equalizers and, 91 local multipoint distribution service and, 318, 320 digital watermarking, 1077 digital wrappers, signal quality monitoring, 2269 digitization, waveform coding, 2830-34 Dijkstra's algorithm, 550, 2208 dimensionality, coding division multiple access, 458-461 dipole antennas, 169, 180 active antennas and, 48 antenna arrays and, 142 crossed drooping, 197, 197 directivity in, 1258 gain in, 1258 impedance, impedance matching in, 1258 linear, 1257-58, 1257 radiation pattern in, 1257-58, 1258 television and FM broadcasting, 2517-36 direct data domain least squares method, 71-73 direct detection, wireless infrared communications, 2926 direct detection receivers, optical transceivers, 1825 direct digital frequency synthesizer, 447 direct matrix inversion, blind multiuser detection, 298, 300, 306 direct sequence CDMA, 458, 459, 1602 adaptive receivers for spread-spectrum system and, 96, 97, 101, 104, 107 Bluetooth and, 310 in channel modeling, estimation, tracking, 409, 411 chaotic systems and, 422, 428, 428 chirp modulation and, 445 cochannel interference and, 455 iterative detection algorithms for, 1196–1210 media access control and, 1345 multicarrier CDMA and, 1522, 1523 multiple input/multiple output systems and, 1456 packet rate adaptive mobile receivers and, 1886-87, 1886 polyphase sequences and, 1975-82 powerline communications and, 2003 radio resource management (RRM) and, 2090, 2091-93 shallow water acoustic networks and, 2209 signature sequence for CDMA and, 2274, 2283, 2284 spatiotemporal signal processing and, 2336 Viterbi algorithms and, 1196-1210 wireless multiuser communications systems and, 1608, 1614, 1615 direct sequence spread spectrum, 1130-41, 2392-96, 2392 acoustic telemetry in, 23, 27 blind multiuser detection and, 298 diversity and, 733-734 interference and, 1130-41 multicarrier CDMA and, 1521, 1523 shallow water acoustic networks and, 2216-17 signature sequence for CDMA and, 2274 in underwater acoustic communications, 41 wireless communications, wireless LAN and, 1285, 2678, 2842–43, 2842 direct sequence ultrawideband radio, 2757 direct sequency FSK, 1474 direct solution algorithm, cyclic coding, 617 direct video broadcast, broadband, 2671-73, 2672

directional antennas, 198, 1348 directivity of antennas, 180, 185, 186, 196 linear antennas and, 1258 loop antennas and, 1293-94 microstrip/microstrip patch antennas and, 1360-1361 active antennas and, 63 directivity gain, in antenna arrays, 142 discrete autoregressive model, traffic modeling, 1666, 1668 discrete cosine transform image and video coding and, 1039 image compression and, 1066, 1067 image processing and, 1074, 1075 orthogonal transmultiplexers and, 1881, 1881 vector quantization and, 2125-26 waveform coding and, 2837 wireless MPEG 4 videocommunications and, 2972-81 discrete Fourier transform adaptive equalizers and, 86 broadband wireless access and, 321 image processing and, 1074, 1075 orthogonal frequency division multiplexing and, 1871-72, 1944 orthogonal transmultiplexers and, 1880-85 simulation and, 2287, 2288 waveform coding and, 2837 discrete Hadamard transform, 2837 discrete logarithm problem, cryptography, 609-610 discrete memoryless channels, information theory, 1113-14 discrete multitone, 736-748, 738 additive white Gaussian noise, 745 applications for, 746-747 asymmetric DSL and, 746-747, 747 baseband and, 741-742 bit loading in, 745 digital signal processing and, 740-741 frequency division multiplexing and, 736-737, 737 frequency domain equalization in, 745 guard interval in, 739-740, 739 matrix notation in, 743-745 orthogonal frequency division multiplexing and, 736, 1878, 1944 quadrature amplitude modulation and, 737 Shannon or channel capacity in, 745-746 signal to noise ratio, 745 spectral properties of, 742-743, 743 transmitters and receivers for, 740, 742 very high speed DSL and, 2791-2801, 2792, 2794 discrete time Fourier transform, digital filters, 690-691 discrete time function, orthogonal transmultiplexers, 1880 discrete time models, traffic modeling, 1667 discrete time signals, digital filters, 687-689 discrete Walsh transform, waveform coding, 2837 discrete waveform/wavelet transform image and video coding and, 1040 image compression and, 1067-68 image processing and, 1074, 1075 underwater acoustic communications, 37 discrete-time, in channel modeling, estimation, tracking, 399-401 discretionary access control, 1649 dispatch radio services, 1478 dispersion fading and, 784 leaky wave antennas and, 1237, 1237 microwave and, 2565 optical fiber and, 436, 1709, 1711, 1712-13 orthogonal frequency division multiplexing and, channel time, 1874 polarization mode (see polarization mode dispersion) solitons and, 1764, 1765 wavelength division multiplexing and, 2869 dispersion compensated optical fiber, 1712-13 dispersion compensating devices, 1848 dispersion compensating fiber, 1686, 1768, 1846 dispersion compensating module, 1484 dispersion management, optical fiber, 1686 dispersion shifted fiber, 1711, 1714, 1845, 1848

dispersive fade margin, microwave, 2565, 2566 distance routing effect algorithm for mobility, 2890 distance vector multicast routing protocol, 1534 distance vectors, 1534 distorted Gaussian models, traffic modeling, 1670 distortion adaptive equalizers and, 79 amplitude modulation and, 134 blind equalizers and, 286 digital audio broadcasting and, 677 digital magnetic recording channel and, 1325-1326 image compression and, 1063 intersymbol interference and, 1157-62, 1158-61 nonlinear (see also RF power amplifiers, nonlinear distortion in) optical communications systems and, 1484-85 orthogonal frequency division multiplexing and, 1876 rate distortion theory and, 2069-80 distortion bounds, compression, 640-641 distortion rate function, 640-641, 2123 distributed admission control, 118 distributed Bellman-Ford algorithm, 2886 distributed Bragg reflector laser, 1780-81, 1780 distributed COM, 725, 727 distributed coordinated function, 1348 distributed feedback laser, 1779, 1779, 1780 free space optics and, 1853-54 optical signal regeneration and, 1762 distributed intelligent networks, 719-29, 722, 726 distributed mesh feedback photonic AD conversion, 1967 distributed protocols, media access control, 5, 1343 distributed queue dual bus (DQDB), medium access control, 1558 distributed queue multiple access, medium access control, 1558 distributed routing, multimedia networks, 1566 distribution, antenna arrays, synthesis by, 154 distribution system, for wireless communications, wireless LAN, 1285 diurnal variations inradiowave propagation, 2063, 2065 divergence factors inradiowave propagation, 211 diverging edge, in trellis coded modulation, 2627 diversity, 371, 729-736 additive white Gaussian noise, 732, 733 applications for, 733-735 bit error rate and, 732, 733 cdma2000 and, 361, 367 channels and, 729 cochannel interference and, 456 combining techniques for, 730-733 cross correlation and, 729 fading and, 730-731, 787-788 gain and, 729 interleaving and, 733 IS95 cellular telephone standard and, 355-356 maximal ratio combining in, 731 microwave and, 2563-69 multipath and, 730-731 multiple input/multiple output systems and, 1455 optical fiber and, 735 optimal selection, 731 outage rates and, 729 quadrature amplitude modulation and, error probability and, 2050-52 radio resource management and, 2093 RAKE receivers and, 732, 734 satellite communications and, 1230-31 scanning, 731 signal to noise ratio, 731 space-time coding and, 2324 spatiotemporal signal processing and, 2333, 2334-36 tropospheric scatter communications and, 2693-95 wireless and, 2919-20 wireless multiuser communications systems and, 1603, 1608 wireless transceivers, multi-antenna and, 1583, 1584 diversity gain, 1450-52, 1451, 1452, 2324 diversity reception, in antennas for mobile communications, 190

diversity techniques, 1481

divisive methods in quantization, 2128 DNA sequence analysis, andViterbi algorithm, 2818 DoCoMo, 392, 2193 DOCSIS (see data over cable service interface specifications), 272 Dolph-Chebyshev linear antenna arrays, 145-146, 147, 187 domain name servers, 548, 1913, 2199 domain convolution, simulation, 2287 dominant mode waveguides, 1390 dominant path concept, indoor propagation models, 2020, **2021** doping for optical fiber, 1484 Doppler effect, Doppler fading, Doppler spreading, 442 in acoustic modems for underwater communications, 18 - 19acoustic telemetry in, 23 cellular communications channels and, 394, 395 in channel modeling, estimation, tracking, 410, 412, 413 chirp modulation and, 441 intelligent transportation systems and, 509-510 mobile radio communications and, 1481 multiple input/multiple output systems and, 1453 power control and, 1983 satellite communications and, 196, 197 shallow water acoustic networks and, 2207 simulation and, 2291 tropospheric scatter communications and, 2698-99 in underwater acoustic communications, 39-40, 39 wireless and, 2916-18, 2917, 2918, 2916 wireless multiuser communications systems and, 1604 Doppler frequency, 783-785, 2325 Doppler power spectrum, wireless multiuser communications systems, 1604 Doppler sensors, active antennas, 65 double random phase encryption, 2132-33 double sideband AM, 133 double sideband suppressed carrier AM, 133-134, 133, 140, 140 doubletalk, acoustic echo cancellation, 4, 11 downlink, in satellite communications, 877, 1223, 1223, 2115 drift ALOHA protocol and, analysis in, 129, 129 frequency synthesizers and, 837 DRIVE project in intelligent transportation systems, 503 drive-response synchronization and chaos, 422, 422 driving point, optical modulators, 1746 dropouts, digital magnetic recording channel, 1326 DropTail algorithms, flow control, 1627, 1628, 1660 DSÂ, 218 dual beam antennas, 191, 194 dual busy tone multiple access, 2210, 2886 dual frequency antennas, 191, 194 dual polarized waveguide antenna, 1416-17, 1417 dual queue dual bus, optical fiber, 1715 dual tone multifrequency, session initiation protocol (SIP), 2198 ducting, millimeter wave propagation, 1435, 1445 Duffy's transform, in antenna modeling, 174 duobinary and modified duobinary signals, 1829–30, 1829 duobinary encoder, minimum shift keying, 1474, 1475 duobinary frequency shift keying, 585 duobinary pulse, partial response signals, 1930–32, **1931** duplexing, very high speed DSL, 2793 DVD-ROM media, 1738 dynamic allocation schemes, medium access control, 1553 dynamic bandwidth allocation, Ethernet, 1511 dynamic channel allocation, radio resource management, 2091-93 dynamic feedback modulatioin, chaotic systems, 422, 423, 423 dynamic host configuration protocol, 335, 869 dynamic mode cell planning in wireless networks, 388, 390-391

dynamic range, in optical fiber, 1718 dynamic restoration, 1635

image and video coding and, 1031-33

image compression and, 1065-66

transform coding and, 2596

dynamic routing, wavelength assignment, 2101-04 electrooptic bottlenecks, routing and wavelength assignment in WDM, 2098 dynamic source routing, 2211, 2888 dynamic time warping, automatic speech recognition, electro-optic planar lightwave circuits, 1703-04 electroptical switches, 1785 2373 dynamic time wavelength division multiple access, 1552 element patterns, antenna arrays, coupling, 164-166, 165, 166 E plane stepped DFW, 1411-16, 1413-16 elementary symmetric functions, 248, 623 earliest available time scheduling, 1554 elements in antenna arrays, 144, 180 Ellipso satellite communication, 196 earliest due date algorithm, flow control, traffic management, 1660 elliptic curves, cryptography, 610, 610, 613 early congestion notification (ECN), multiprotocol label email, 540 switching, 1594 embedded coding, speech coding/synthesis, 2354-55 embedded Markov chains, traffic modeling, 1668 early late gate synchronizer, pulse amplitude modulation, 2029, 2029, 2057 embedded radio systems, Bluetooth, 309 early packet discard, flow control, traffic management, embedded wavelet coding, image compression, 1069-70 1661 emergency services, session initiation protocol (SIP), 2203-04 earth bulge, microwave, 2559 Earth Observation Satellite waveguides, 1391, 1392 emission earth reflection, 209-210, 209 lasers and, 1776 Earth-space transmission paths, millimeter wave propamillimeter wave propagation and, 1436-37, 1445-46 gation, 1445 EMMA probe, shallow water acoustic networks, 2206 empirical path loss prediction models, radiowave propaeavesdropping, 2810 Ebert-Fastie spectrograph and diffraction gratings, gation, 216 1751, **1752**, 1754-55 empty cluster problem in quantization, 2129 encapsulating security payload, virtual private networks, echo, in IP telephony, 1178 2810-11, 2810, 2811 echo cancellation filter, 1-8, 2, 12 echo cancellation, acoustic (see acoustic echo cancellaencapsulation frame, 542, 542 tion) echo model, powerline communications, 2000-2001, multiprotocol label switching and, of labels, 1594, 1594, 1597-98 2001 virtual private networks and, 2708-09 echo return loss enhancement, acoustic echo cancellation, 3 encapsulation, of messages, 540 ECHO satellite communication, 196 encipherment (see also cryptography), 1648-49 economies of scale, traffic engineering, 494 encoders/encoding edge networks, optical fiber systems, 1840, 1841 in BCH (nonbinary) and Reed-Solomon coding, 468-469 edge routers, burst switching networks, 1802 CDROM and, 1734 effective area, antennas, 180, 186 effective bandwidth, 809, 1671, 1673, 1908 CIRC encoders for, 627-628, 627 effective index, lasers, 1777 constrained coding techniques for data storage and, effective isotropic radiated power, 881, 1268 573, 575-576, 575 effective length, in optical fiber, 1489 convolutional coding and, 599-600, 599 effective number of bits, in cable modems, 327 cyclic coding and, 619-620 effective spreading coding, in adaptive receivers for linear predictive coding and, 1264 spread-spectrum system, 106 low density parity check coding and, 1316-17 efficiency millimeter wave antennas and, 1425 magnetic storage and, 1326-1333, 1327 parabolic and reflector antennas and, 1923-24 modems and, 1497 multiple input/multiple output systems and, 1455-56 efficient reservation virtual circuit, 552-553, 552 EFR algorithm, speech coding/synthesis, 2827 optical synchronous CDMA systems and, 1809 egress edge router, burst switching networks, 1803, 1803 orthogonal frequency division multiplexing and, egress routers, differentiated services, 668 1873, 1876 product coding and, 2007, 2009 Eigen algorithm, interference, 1116-19 sequential decoding of convolutional coding and, eight to fourteen modulation, 579, 1735 2141, 2141 El Gamal encryption, 612, 613, 1649 trellis coding and, 2635 elastic sources in traffic modeling, 1671, 1672-73 electrabsorption modulators, 1770, 1771 turbo coding and, 2705, 2705 ultrawideband radio and, 2755-58 electric field integral equation, 173 in underwater acoustic communications, 43, 45-46 electric fields vector quantization and, 2125 active antennas and, 61 antennas, 180 waveform coding and, 2834-37 waveguides and, 1394 encryption (see cryptography) electrical equivalents, 180 end systems, 549 end to end connections, 539 electrical power lines (see powerline communications) end to end delay, flow control, 1627 electro absorption modulated lasers, 1826 electroabsorption modulators, 1761-62, 1826 endfire antenna arrays, 142, 145, 148, 149 endpoint admission control, 118 electroacoustic transducers (see acoustic transducers) electromagnetic compatibility, 1995-96, 2001 energy bands in lasers, 1777 electromagnetic interference, 1390 energy consumption (see power management) energy density in active antennas, 54 electromagnetic spectrum, 1423-25, 1424 energy function, neural networks, 1677 electromagnetic theory, 208 electromagnetic wave mode propagation, waveguides, enhanced data rate for global evolution, 126, 369, 383-385, 385, 908, 1096-1108, 2589 1390 electronic cash, 615 enhanced full rate coders, speech synthesis/coding, 1306 electronic codebook, 607 enhanced variable rate coder, 1306, 2827 Electronic Industries Alliance, 434 ENIAC, 2653 electronic serial numbers, IS95 cellular telephone stanensemble-averaged autocorrelation, 428 enterprise networking, broadband, 2656-66 dard, 355 electronics noise, optical transceivers, 1833 enterprise system connectivity, 2865 entire-domain basis functions, antenna modeling, 174 electrooptic (Pockels) effect, optical modulators, 1742 electrooptic analog to digital converters, 1961-64, 1962, entropy bounds, compression, 633-634

1963

entropy coding

entropy constrained vector quantization, 2128 ENUM mechanism in session initiation protocol, 2198 enumerative coding, compression, 635-636 envelope detectors, amplitude modulation, 134-135, 135, 139–140, 139 envelope functions, pulse amplitude modulation, 2022 envelope power function, peak to average power ratio, 1948 envelope processes, traffic modeling, 1666, 1673 equal cost multipath, multiprotocol label switching, 1599 equal gain combining, wireless, 1527, 2920, 2921-22 equalization/equalizers, 79-94 acoustic telemetry in, 26 adaptation algorithm in, 82 adaptive (see adaptive equalizers) baud vs. fractional rate, 286 Beneveniste-Goursat algorithm in, 92 blind (see also blind equalizers), 79, 82, 91-93, 286-298, 1680 in channel modeling, estimation, tracking, 417, 417 channel estimator in, 90 classification of, and algorithms used in, 81-82, 81 constant-modulus algorithm in, 92 decision feedback (see decision feedback equalizers) decision-directed mode in, 82 digital video broadcasting and, 91 distortion and, 286 fast startup equalization in, 82 feedforward, 330 filters in, 81-82, 89 finite impulse response transversal filter in, 81 fractionally spaced, 288–289, **289** infinite length symbol spaced, 1690-92 intersymbol interference and, 80-81, 87, 286, 291 concatenated convolutional coding and, 556-558, 556 iterative least squares with enumeration, 292 lattice filter in. 81 least mean square algorithm in, 83-84, 85, 88, 90, 286 least squares, 81-82, 84-85 linear adaptive (see also adaptive equalizers), 82-87, 82 linear, 286 M algorithm in, 81 magnetic recording systems and, 2258-63 map symbol-by-map symbol, 89, 89 maximum a posteriori in, 81, 89 maximum likelihood sequence estimation, 286 maximum likelihood, 89-91 mean squared error, 81, 83-84, 84, 86, 88 microwave and, 2565, 2569-70 midamble in. 90 minimax criterion in, 81, 82-83 minimum mean square error, 292, 321 multiple input/multiple output, 93 neural networks and, 1679-80 noniterative algorithms in, 82 Nyquist theorem and, 86 orthogonal frequency division multiplexing and, 93 pseudonoise in, 85 recursive least squares algorithm in, 82, 84-85, 85, 90 recursive mean squares, 286 reference signals and, 85-86 sampling in, 286, 287 Sato algorithm in, 92 signal to noise ratio in, 88 single carrier frequency domain equalization in, 2329–30, **2330** space-time coding and, 2327-30 spacing in, symbol-spaced vs. fractionally spaced, 86-87, **87** startup equalization in, 81 stop-and-go algorithm in, 92 system model for, 79-80, 79, 80 tap leakage algorithm in, 87 tapped delay line, for sparse multipath channels, 1688-96

equalization/equalizers (continued) training mode in, 82 training sequences in, 286, 287 trellis-coded modulation and, 91, 91 turbo type, 2716-27 in underwater acoustic communications, 42-43 underwater communications and, 16 Viterbi algorithm in, 81, 90 whitened matched filter in, 89 zero forcing, 81, 82-83, 88 equation error, blind equalizers, 293 equilibrium, flow control, 1629-30 equipment identity register, global system for mobile, 906 equivalent circuits, active antennas, 29, 60-61, 61 equivalent encoders, convolutional coding, 599-600 equivalent noise current density, optical transceivers, 1833 erasure filling decoding, BCH coding, 247, 251-252, 259-261 erasure probability, sequential decoding of convolutional coding, 2159 erbium doped fiber amplifier, 706, 1484, 1835 Bragg gratings in, 1728 free space optics and, 1853 Gigabit Ethernet and, 1509 optical crossconnects:, 1702 optical fiber and, 1709, 1721, 1842 optical multiplexing and demultiplexing and, 1748 signal quality monitoring and, 2273 solitons and, 1764 wavelength division multiplexing and, 2839, 2869 ergodicity, in wireless multiuser communications systems, 1606 Erlang B blocking capacity and, 492, 494 cell planning in wireless networks and, 379-380, 379 380 cochannel interference and, 453, 453, 454 economies of scale in, 494 flow conservation principlie in, 493-494, 493 Markov chains in, 492 seizure of a server by call in, 492 service facility in, 492 state space in, 492-493, 493 state transition diagrams in, 492-493, 493 tables for, 494 traffic engineering and, 487, 491-499, 498 trunking efficiencies in, 494 Erlang C blocking, 495-497, 495 Erlangs, traffic engineering, 486 error control/correction coding constrained coding techniques for data storage and, 570, 579 convolutional coding and, 598 cyclic coding and, 616-630 low density parity check coding and, 1316 magnetic storage and, 1326 multidimensional coding and, 1539-40, 1539 partial response signals and, 1933 Reed-Solomon coding for magnetic recording channels and, 466-467, 466, 470, 472-474 threshold decoding and, 2579–85 trellis coding and, 2635, 2639–40 turbo coding and, 2704-05, 2716-27 underw/in underwater acoustic communications, 40-41 unequal error protection coding and, 2762-69 wireless infrared communications and, 2927-28 error criterion, adaptive antenna arrays, 72-73 error detecting coding, Reed-Solomon coding for magnetic recording channels, separate vs. embedded, 474 error detection and correction, 545, 1633 ATM and, 2908-09 automatic repeat request and, 224-231 BCH coding, binary, and, 238-253 BCH/ in BCH (nonbinary) and Reed-Solomon coding, 253-262 Bluetooth and, 313 continuous phase frequency shift keying and, 594-598, **595**, **596**

continuous phase modulation and, 587-590, 2182-84, 2183, 2182 cyclic coding and, 620 image and video coding and, 1030-31 IS95 cellular telephone standard and, 350, 354 magnetic recording systems and, 2256-58, 2257 magnetic storage and, 4, 1332 multidimensional coding and, 1540-41, 1544-48 orthogonal frequency division multiplexing and, 1875-76 packet rate adaptive mobile receivers and, 1886 permutation coding and, 1955-56 powerline communications and, 2002, 2004 pulse amplitude modulation and, 2022, 2024-25, 2027 pulse position modulation and, 2036-39 quadrature amplitude modulation and, 2047-50 satellite communications and, 1223, 1229-31, 1230, 1231 sequential decoding of convolutional coding and, 2159-60 serially concatenated coding and, 2165-66, 2166, 2182-84, 2183 shallow water acoustic networks and, 2207 sigma delta converters and, 2229-30, 2229 signal quality monitoring and, 2269 speech coding/synthesis and, 2342, 2355, 2367 spread spectrum and, 2394-95 trellis coding and, 2639-40 tropospheric scatter communications and, 2701-02 unequal error protection coding and, 2762-69 wireless IP telephony and, 2936 wireless MPEG 4 videocommunications and, 2978 error floor region, serially concatenated coding, 2166-67, 2167 error locators, BCH coding, 623 error rate definitions, Reed–Solomon coding for magnetic recording channels, 473 error resilient entropy coding, 1576 error trapping decoder, 247, 251, 617 estimation theory, 0, 1338 etching, microelectromechanical systems, deep reactive ion, 1352 Ethernet, 549, 1280–1281, 1501–13 10Base2, 1283 10Base5, 1283 10BaseT, 1283, 1283 100BaseT, 1283-84 addressing and, 1503 architecture for, 1502-05, 1502 asynchronous transfer mode (ATM) and, 1512 attachment unit interface in, 1506 bridges in, 1505 broadband and, 2655 broadcast domains in, 1281 burst mode in, 1284 bus architecture in, 1503, 1504 carrier sense multiple access and, 345, 1280-81 carrier sense multiple access with collision detection and, 1503-04, 1504, 1505 coaxial cable for, 1283, 1506-07 collision domains in, 1281 collisions in, 1280 copper PHY with extended reach and temperature, in EFM, 1510 cyclic redundancy check in, 1503 development of, 1501, 1502 dynamic bandwidth allocation in, 1511 Ethernet in the First Mile in, 1289, 1508-12, 1510, 2803-05, 2804 failure and fault detection/recovery in, 1633-34 fiber distributed data interface in, 1284 free space optics and, 1851 full duplex, 1284 Gigabit, 1284, 1501, 1507, 1508-09, 1510 half duplex operation in, 1504 hubbed architectures for, 1505, 1505 jamming in, 1281, 1283 layers of, 1502-05, 1502 link aggregation in, 1284 local area networks and, 1279, 1280-81, 1501, 1512

MAC frames in, 1502-03, 1503 MAC layer in, 1502 media access control and, 9, 1347, 1506 medium access unit in, 1506 medium dependent interface in, 1508, 1509 medium independent interface and, 1502, 1506, 1507 metropolitan area networks and, 1512 multicasting and, 1529-30 optical fiber and, 1507, 1510-11, 1717, 1719 optical line termination in, 1511, 1512 optical network unit in, 1511, 1512 OSI reference model and, 1502-05, 1502 packet switched networks and, 1910 passive optical networks and, 1510-12, 1511, 1512 physical coding sublayer in, 1507-08 physical layer in, 1502, 1506-08 physical medium attachment sublayer in, 1508 physical medium dependent sublayer in, 1508 point to multipoint operation, in EFM, 1511-12 pulse amplitude modulation and, 1508 repeaters in, 1504-05 security and, 1646 session initiation protocol and, 2197 signal quality monitoring and, 2269 64B/66B encoding in, 1508 slot time in, 1281 SONET vs., 1501, 1512 source address table in, 1505-06 start frame delimiter in, 1503 switches in, 1505-06, 1506 thinnet/cheapernet, 1283 time division multiplexing and, 1512 topologies for, 1505, 1505 transmission media for, 1506-07 truncated binary exponential backoff in, 1281 unshielded twisted pair in, 1283, 1506-07 virtual LAN and, 1284 wavelength division multiplexing in, 1507 wide area networks and, 1512 Wireless Ethernet Compatibility Alliance and, 1288 wireless LAN and, 1284-89 Ethernet for the First Mile, 1289, 1509-12, 1510, 2803-05, 2804 ETS-V satellite communications, 198 Euclid's algorithm, cyclic coding, 617 Euclidean distance low density parity check coding and, 661-662 magnetic recording systems and, 2249, 2260 pulse amplitude modulation and, 2025 serially concatenated coding and, 2173 serially concatenated coding for CPM and, 2182 speech coding/synthesis and, 2355 trellis coded modulation and, 2622, 2627-29 trellis coding and, 2642 Euclidean geometry coding, 802-807 Euler algorithms, chaotic systems, 424 European Advanced Communications Technologies and Services, 1720 European mobile satellite, 2112 European Telecommunications Standards Institute (see standards) evaluation function, sequential decoding of convolutional coding, 2145 evanescent waves, waveguides, 1395 even parity, multidimensional coding, 1540 event detection, 1650 excess loss, optical couplers, 1699 excitation speech coding/synthesis and, 2341, 2347-48 waveguides and, modes of, 1397-1405, 1398 exclusive OR gates, signal quality monitoring, 2270 existential forgery attacks, 612 exit charts, serially concatenated coding, 2167, 2172 expansion coefficients in antenna modeling, 174, 176-177 expectation maximization algorithm, 769-780 blind equalizers and, 290 channel estimation and, 771-772 fading and, 772, 776-778, 778 interference channel and, 772-773 maximum likelihood estimation and, 3, 1341

expectation maximization algorithm (continued) multiuser channel estimation and, 771 nonconvergence in, 774-778 orthogonal frequency division multiplexing and, 773 parameter estimation from superimposed signals and, 770-771 random phase channels and, 772 signal to noise ratio, 773-774, 774, 775 space time coding and, 773 unsynchronized channels and, 772 expedited forwarding, DiffServ, 271, 669-670, 673-675, 1657-58 explicit cell rate, ATM, 552 explicit congestion notification, flow control, traffic management, 1662-63, 1662 explicit forward congestion indication, ATM, 200, 206 explicit rate feedback, flow control, traffic management, 1663 explicit rate indication for congestion avoidance, congestion control, 1663 explict congestion notification, flow control, 1628 explict routed LSP, multiprotocol label switching, 1592, 1593, 1593 exponential backoff, carrier sense multiple access, 345 exponential filtering, in channel modeling, estimation, tracking, alpha trackers, 415 exponentially windowed RLS, in channel modeling, estimation, tracking, 415 exposed terminal problems, 5, 1343, 2885, 2885 extended BCH coding, binary, 246 extended partial response, 4, 1328-1331, 1331 extended service set, wireless LAN, 1285, 1286 extending Reed-Solomon coding, 467 extensible HTML, wireless application protocol, 2899 extensible markup language, 1651 distributed intelligent networks and, 728 software radio and, 2304 wireless application protocol and, 2899 extension fields, BCH coding, binary, 240-241 exterior gateway protocol, 269, 1913 exterior router protocols, 549 external network to network interface, 1799 extinction ratio, 2577, 2577 optical fiber systems and, 1842 optical modulators and, 1743 optical signal regeneration and, 1760 extranets, 1163-72, 1165 extremely low frequency (see also atmospheric radiowave propagation), 758–780 eye patterns, 2576–79, **2576**–2578 eye safety and lasers, 1864-65 fabrication attacks, 1151 Fabry-Perot interferometer, 1003, 1723-25, 1723, 1724, 1725, 1749, 1756-57, 1757 facet loss, in lasers and, 1778-79 factor bases, in cryptography and, 610 factor graphs, for low density parity check coding and, 1316 factors of merit (see merit factor) fading (see also multipath), 208, 781-802, 2065, 2066 in acoustic modems for underwater communications, 15 additive white Gaussian noise, 786-787 antennas for mobile communications and, 190 autocorrelation and, 783 bit error rate (BER) in, 787 bit interleaved coded modulation and, 276, 278, 280, 281, 283, 285 cellular communications channels and, 393, 394 chann/in channel modeling, estimation, tracking, 410, 410 chaotic systems and, 430-431, 430 chirp modulation and, 446 cochannel interference and, 449 delay and, 783 dispersion and, 784 diversity and, 730-731, 787-788 Doppler frequency shift and, 783, 784, 785 expectation maximization algorithm and, 772, 776-778, 778

free space optics and, 1862-63 high frequency communications and, 949 indoor propagation models and, 2013, 2013 large scale, 781-782 location in wireless systems and, 2965, 2967 maximal ratio combining and, 788 mean square error and, 781 microwave and, 2562-65, 2562, 2571 mobile radio communications and, 1481 multiple input/multiple output systems and, 1455-54, 1454 packet rate adaptive mobile receivers and, 1886 path loss and, 1937 power control and, 1983 quadrature amplitude modulation and, 2050-52 RAKE receivers and, 787-788 satellite communications and, 1223, 1226-29, 1226, 1227 shadowing and, 781 signal characteristics and, 784-785 signal to noise ratio, 786-788 simulation and, 2290-91, 2291 space-time coding and, 2324 spatiotemporal signal processing and, 2333-40, 2333 spectral broadening and, 784 statistical characteristics of, 783 trellis coded modulation and, 2633-34 tropospheric scatter communications and, 2698-99 underw/in underwater acoustic communications, 40, 45 Viterbi algorithm and, 2817-18, 2817 wireless and, 2915, 2916-18, 2916 wireless multiuser communications systems and, 1603-11 wireless transceivers, multi-antenna and, 1579 fail stops, 1632 failure and fault detection/recovery in, 1632-34 failures, 1631 fair distributed queue, medium access control and, 1558 fairness, flow control and, 1626, 1653, 2103 FairNet in medium access control and, 1558 fan antennas, 180 Fano algorithm, Fano metric, convolutional coding and, 2140, 2146–48, 2150–54, 2151, 2152, 2153, 2154 fanout of power splitters, 2104 far end crosstalk, 2786, 2798–2800, 2803–05, 2805 far field (Fraunhofer region) antenna arrays and, 141 antennas, 181-182, 181-182, 182 loop antennas and, 1292, 1293 multibeam phased arrays and, 1514 parabolic and reflector antennas and, radiation concepts in, 2080-81, 2080 Faraday law in antennas and, 171 Faraday, Michael, 208 Farley's approximation, cochannel interference and, 451 fast algorithms, antenna modeling and, 173 fast broadcasting, 236 fast fading path loss and, 1937 simulation and, 2290-91, 2290 wireless multiuser communications systems and, 1605 fast Fourier transform adaptive antenna arrays and, 72 antenna modeling and, 173 location in wireless systems and, 2970 multicarrier CDMA and, 1522 orthogonal frequency division multiplexing and, 1871 signal quality monitoring and, 2272 simulation and, 2288 fast frequency shift keying, 584 fast multipole method, for antenna modeling and, 173 fast resource management, ATM and, 552 fast startup equalization, 82 fatigue testing, optical fiber and, 438-439 FatMAC protocol, 1553 fault isolation, signal quality monitoring and, 2269 fault management, ATM and, 206-207 fault tolerance, 1631-44 asynchronous transfer mode and, 1633-35 automatic repeat request and, 1632

backup schemes and, 1634-35 circuit switched networks and, 1632 cycle covers and, 1638-39, 1638 cyclic redundancy check in, 1633 dynamic restoration in, 1635 fail stops in, 1632 failure and fault detection/recovery in, 1632, 1633-34 failures and, 1631 fault isolation boundaries in, 1632 fiber distributed data interface and, 1637 intermittent failures and, 1631 link and node-based schemes for, 1635 link rerouting in, 1633-34, 1634 Menger's theorem and, 1635 mesh networks and, 1637-39, 1638, 1639 metropolitan area networks and, 1632 minimum spanning tree in, 1639-40 models for, 1632 multiprotocol label switching and, 1640 optical fiber and, 1636, 1636 packet switched networks and, 1632, 1639-40 path and link monitoring in, 1633 path-based schemes for, 1634-35, 1634 protection of links or nodes in, 1634 quality of service and, 1632 redundancy and, 1632 rings for, 1635–37, **1636** self-healing rings in, 1635, 1637, 1638 SONET and, 1634, 1635 subnetwork connection protection and, 1635 topologies for, 1632-33 transmission control protocol and, 1632, 1640 wide area networks and, 1632 fax modems, 1499 FCC clear channel skywave curve in, 2061-62 FEC to NHLFE, multiprotocol label switching and, 1594 Federal Communications Commission, 309 Federal Information Processing Standards, cryptography and 606 feedback flow control and, 1626 frequency modulation (FM) and, 814-815 lasers and, 1776-77 synchronization and, 2475 feedback (return) channel, automatic repeat request and, 224 feedback algorithms, speech coding/synthesis and, 2354 feedback circuits, active antennas and, 51, 51 feedback control, optical modulators and, 1746 feedback filters, tapped delay line equalizers and, 1690 feedback loops, sigma delta converters and, 2230, 2232, 2233–47 feedback networks, active antennas and, 58-59, 59 feedback shift register and FSR synthesis, 257-259, 789-802 autocorrelation in, 795, 798, 799 balance property in, 795 Berlekamp-Massey algorithm in, 790, 797-798 code division multiple access, 789 complemented cycling registers in, 799 complemented summing registers in, 799 constant on the coset property in, 795–796 cross correlation in, 796 cycle and add property in, 796 cyclic Hadamard difference sets in, 790, 795 De Bruijin sequences in, 790, 795 disjoint cycles and, 798-799 linear, 790 linear recurring sequences in, 790 m sequences in, basics of, 791-795, 791 primitives in, 792 pseudonoise sequences and, 789 pure cycling register, 794, 794, 796-797, 798 pure summing register in, 794, 794, 799 run distribution property in, 795 span property in, 795 spread spectrum and, 789 state diagrams for, 790-791, 790, 793, 794 trace function in, 796 truth tables for, 790-791, 790, 793, 794

feedback systems, neural networks and in, 1676-77

feedback/feedforward encoders, convolutional coding and, 599-600, 599 feeder links, satellite communications and, 1251 feedforward synchronization and, 2475 feedforward algorithms, in speech coding/synthesis and, 2354 feedforward equalizers, 330, 1973-74, 1973 feedforward filters, tapped delay line equalizers and, 1690 feedforward vs. feedback systems, neural networks and in, 1676-77 feeds antenna arrays and, 166, 166, 167 horn antennas and, 1006-17, 1006-16 leaky wave antennas and, 1245 microstrip/microstrip patch antennas and, 1361-1363, 1373–1374, **1373**, **1374**, 1380, 1383–1384, **1383** parabolic and reflector antennas and, 1920, 1924, 2082, 2083, 2084 waveguides and, 1392-1393, 1392 Fekete's lemma, constrained coding techniques for data storage and, 573, 574 Fenton-Wilkinson approximation, cochannel interference and, 450, 451 ferrite loaded loop antenna, 1296–97, 1296 Fiat Shamir identification protocol, authentication and, 614 fiber delay lines, burst switching networks and, 1804-06, 1805, 1804 fiber distributed data interface, 547, 1284, 2461 carrier sense multiple access and, 345 fault tolerance and, 1637 free space optics and, 1851 media access control and, 8, 1346 optical crossconnects:, 1701 optical fiber and, 1715, 1718–19, 1808 reliability and, 1637 wavelength division multiplexing and, 2864 fiber optic test procedures, 434 fiber optics (see optical fiber systems) fiber ring lasers, solitons and, 1771 fiber stress history, 439 fiber switch capable interfaces, 1799 fiber to the building and, 1797 fiber to the curb, 1797, 2957 fiber to the home, 1797, 1808, 2957 Fibre Channel, 1641, 1719 field effect transistors, 51, 57-58, 57, 58 field equivalence principle, antennas, 183-184 field of view, parabolic and reflector antennas and, 1924 field programmable gate arrays (FPGA), software radio and, 2307, 2316 field regions, antenna, 181–182, **182** field strength, 2064–2067 fields, in BCH coding, binary, and, 238-239 figure of merit antennas, 180, 184-186 optical filters and, 1731-32 optical signal regeneration and, 1759-60 satellite communications and, 1229 file transfer protocol, 540, 541, 544, 1152, 1651 file transfer/data transfer, 1233, **1233**, 1497 filters, 1478 acoustic echo cancellation and, 1-8, 2, 3, 8 acoustoopic, 1729-30 adaptive equalizers and, 81-82 adaptive receivers for spread-spectrum system and, 103-104 amplitude modulation and, 134, 135-136 blind equalizers and, 288-289 blind multiuser detection and, 299 cable modems and, 324-325, 328-329, 333 carrierless amplitude phase modulation and, 336-339 chann/in channel modeling, estimation, tracking, 412-414 characteristics of, 1731-32 chirp modulation and, 442-443, 446, 447 cochannel interference and, 454 continuous phase frequency shift keying and, 596 continuous phase modulation and, 592

control mechanisms for, 1724 corrective, 1723 deemphasis, 821-823 demodulation and, 7, 1335 digital, 686-702 equalizers and, 81-82 equalizers and, 89 image processing and, 1074 lattice, 81 magnetic recording systems and, 2262 magnetic storage and, 1329-1330, 1333 matched, 1335-1338, 1336 optical signal regeneration and, 1764 optical, 1722-33, 1756-58 orthogonal frequency division multiplexing and, 1871 - 72orthogonal transmultiplexers and, 1882-83, 1882, 1884 packet rate adaptive mobile receivers and, 1888–1900, 1893 partial response signals and, 1928 preemphasis, 821-823 pulse amplitude modulation and, 2026, 2029 quadrature amplitude modulation and, 2046, 2049-50 selective, 1723 shallow water acoustic networks and, 2207 sigma delta converters and, 2228, 2232-35 signal quality monitoring and, 2272 signature sequence for CDMA and, 2275 speech coding/synthesis and, 2344-45, 2370, 2378 surface acoustic wave, 2441-61 tapped delay line equalizers and, 1690 tunable, 1724 waveguides and as, 1390, 1416-17, 1417 wavelength division multiplexing and, 2869 wavelets and, 2852-54, 2852 wideband CDMA and, 2878 wireless multiuser communications systems and, 1616 financial cost functions, cell planning in wireless networks and, 374 fine tuner of codebooks in quantization and, 2129 finesse, in optical filters and, 1724 finite antenna arrays and, 165-166 finite element method of antenna modeling and, 170, 176–177, **176** finite fields in BCH coding, binary, and, 238-239 finite geometry coding, 802-807 finite impulse response filters, 693-694, 694, 696-697 adaptive equalizers and, 81 blind equalizers and, 287, 292 digital magnetic recording channel and, 1324 IS95 cellular telephone standard and, 350 magnetic recording systems and, 2259, 2262 magnetic storage and, 1, 1329 random number generation and, 2292-93 sigma delta converters and, 2228 simulation and, 2287-88, 2292-93 speech coding/synthesis and, 2343, 2346 surface acoustic wave filters and, 2450-52, 2456 finite length, in tapped delay line equalizers and, 1692-94, 1694 finite local coanticipation, constrained coding techniques for data storage and, 575 finite state, vector quantization and, 2127 finite state machine, trellis coding and, 2640-42, 2641 finite state transition diagram constrained coding techniques for data storage and, 573 magnetic recording systems and, 2253-57, 2256 finite traceback Viterbi decoding, 602, 602 finite type constraints, constrained coding techniques for data storage and, 571 finite-element boundary integral methods, 170, 177 finline transition, in waveguides and, 1399-1400, 1400 FIR filters, 3, 3, 410 firewalls, 1650-51, 2809 First software radio and, 2305, 2316 first come first served, 234-235, 1565 first fit routing, 2102 first generation wireless systems, 2, 1350

first in first out, 331-332, 487, 495, 1564, 1565, 1627, 1660, 2424 first zone output components, simulation and, 2289 first-generation wireless systems, 370 Fisher's information matrix, maximum likelihood estimation and, 1339, 1340 fisheye state routing, ad hoc wireless networks and, 2889-90 fitting error in blind equalizers and, 293 fixed broadband and, 2671 fixed length prinicipal state coding, 577 fixed priority oriented demand assignment, 9, 1347 fixed rate coding, in scalar quantization and, 2123 fixed satellite services, 877, 1251, 2656 fixed tuned devices, wavelength division multiplexing and, 2840-41 fixed-alternate routing, routing and wavelength assignment in WDM and, 2102 fixed-beam planar, microstrip/microstrip patch antennas and array of, 1386-1387, 1386, 1387 flat earth approximation, in radiowave propagation and, 209 flat fading, 784 spatiotemporal signal processing and, 2334-36 wireless and, 2918-19 wireless multiuser communications systems and, 1604 flat routing protocols, ad hoc wireless networks and, 2889 flatplate slot, antenna arrays and, 142 Fleetsat EHF Packages, 483-484 flexural air ultrasonic transducer, electrorestrictive ceramic, 35, 35 floor acquisition multiple access, 2886 flooring effect, in concatenated convolutional coding and, 560 flow conservation principle, 493-494, 493 flow control (see also congestion avoidance and control; traffic engineering), 545, 1625-31 additive increase multiplicative decrease in, 1630 admission control and, 1625 asynchronous transfer mode and, 1625 ATM and, 550 circuit switched networks and, 1625 complementary slackness in, 1629 congestion control and, 1625, 1627 connection admission control and, 1625 costing, price of links in, 1628-29 delay and, 1627 design objectives in, 1626 DropTail algorithms in, 1627, 1628, 1629 duality model for, 1628-29 dynamic properties and, 1626 end to end delay in, 1627 equilibrium and stability in, 1629-30 explict congestion notification and, 1628 fairness in, 1626 feedback and, 1626 first in first out in, 1627 frequency division multiplexing and, 1625 frequency slots in, 1625 general source/link algorithm in, 1628-29 information constraints to, 1627 modems and, 1496-97 packet switched networks and, 1625, 1911-12 quality of service and, 1625, 1626 queues in. 1626, 1627 random early detection algorithm in, 1627, 1628, 1630 round trip time and, 1627 scalability and, 1626 statistical multiplexing in, 1625 TCP Reno in (see also transmission control protocol), 1625, 1628, 1630, 1662 TCP Tahoe in (see also transmission control protocol), 1625, 1628, 1662 TCP Vegas in (see also transmission control protocol), 1625, 1627-30, 1662 time division multiplexing and, 1625 time slots in, 1625 transmission control protocol and, 1625, 1627-28

flow control (see also congestion avoidance and control; traffic engineering) (continued) transport protocols for optical networks and, 2616-17 utilization and, 1626 window size in, 1627 flows, flow control, traffic management and, 1653 fluid buffer models, in statistical multiplexing and, 2427 - 28fluid traffic models, 1670-71 fluorescent discs in optical memories and, 1739 flutter fading, 2065 flying target algorithm, 1557 focal axis, of parabolic and reflector antennas and, 1920 focal curves of diffraction gratings and, 1751 focal length of parabolic and reflector antennas and, 1920, 2084 focused search technique, in speech synthesis/coding and, 1306 follower sets, in constrained coding techniques for data storage and, 575 footprint of satellite communications and, 1249, 2111 forced erasure decoding, in BCH coding, binary, and, 252 forgetting factor, 8, 9, 101 forking, in session initiation protocol and, 2198 form factors, for hard disk drives and, 1320, 1322 formants, in speech coding/synthesis and, 2361, 2820 forward acknowledgement, congestion control and, 1662 forward equivalence class, 116, 1591 forward error control/correction, 224, 545 automatic repeat request and, 230-231 Bluetooth and, 313 cable modems and, 327, 332-333 cdma2000 and, 359, 360-363 community antenna TV and, digital video in, 524, 525-527, 525 interleaving in, 526 modems and, 1497 optical fiber systems and, 1848 packet rate adaptive mobile receivers and, 1887, 1902 polarization mode dispersion and vs., 1971-72 powerline communications and, 2002, 2004 randomization in, 526 Reed-Solomon coding in, 526 satellite communications and, 878, 1223, 1229-31, 1230, 1231 signal quality monitoring and, 2269 soft output decoding algorithms and, 2295-96 trellis coding and, 526, 2635-40 Universal Mobile Telecommunications System and, 386 wireless IP telephony and, 2933 wireless MPEG 4 videocommunications and, 2973 wireless multiuser communications systems and, 1609 forward fundamental/supplemental coding channels, 356, 359-362 forward link channels, 349-357, 349, 359-362, 361, 367 forward/reverse path, in satellite communications and, 1223, 1223 forwarding differentiated services and, 669-673, 1657-58 IP networks and, 269, 1591 multicasting and, 1532 multimedia networks and, 1568 multiprotocol label switching and, 1591, 1593-95 packet switched networks and, 1909-10 paging and registration in, 1914-15 virtual private networks and, 2808 four photon mixing, 1697-88, 1687, 1712 four wave mixing optical fiber systems and, 1490, 1843, 1846 solitons and, intrachannel, 1769-70 wavelength division multiplexing and, 756 Fourier transforms antenna arrays and, orthogonal method in, 157-158, 158 in BCH (nonbinary) and Reed-Solomon coding, 261 orthogonal frequency division multiplexing and, 1869-70

fourth generation wireless systems, 2, 371–372, 391–392, 1350

four-third's earth radius concept, in radiowave propagation and, 210 fractal antenna arrays and, 142 fractal Brownian motion models, in traffic modeling and, 1669-70 fractal compression, image and video coding and, 1044 fractal Gaussian noise models, traffic modeling and, 1669 - 70fractal images, compression and, 648 fractal Levy motion models, traffic modeling and, 1670 fractional Brownian motion process, 431 fractional N division synthesizers, frequency synthesizers and, 830, 833, 833, 845, 854-862, 860, 862 fractionally spaced equalizer, blind equalizers and, 288-289, 289 fragmentation, in wireless LAN and, 1287 frame in synchronization and, 2482-83 frame check sequence, 547 frame erasure rate, power control and, 1983, 1984 frame error rate, automatic repeat request and, 228-230, 229 frame rate, image and video coding and, 1027 frame relay, broadband and, 2658-59, 2659 frame structure, automatic repeat request and, 225-226, 225 frames, 340, 539, 542, **542**, 545–546 constrained coding techniques for data storage and, 576 encapsulation of, 542, 542 framing, 545-546, 2617 Frank sequence, polyphase sequences and, 1976 Frank-Zadoff-Chu sequence, polyphase sequences and, 1978 Fraunhofer region antennas, 181-182, 181-182, 182 loop antennas and, 1292 multibeam phased arrays and, 1514 parabolic and reflector antennas and, 2080-81, 2080 free distance bit interleaved coded modulation and, 279 continuous phase modulation and, 589 convolutional coding and, 598, 602-604, 603 sequential decoding of convolutional coding and, 2142 serially concatenated coding and, 2167 free space distance, in trellis coded modulation and, 2527-29 free space gratings, 1754-55 free space loss, microwave and, 2556 free space optics, 1849–67, **1851** absorption and, 1851, 1855-57, 1856 atmospheric attenuation in, 1855-57, 1856, 1857 atmospheric refractive turbulence vs., 1861-63, 1861 autocorrelation in, 1862 avalanche photodiode detectors in, 1857 background limited infrared performance in, 1858 bandwidth and, 1849-1850 beamshaping in, 1851 bit error rate in, 1859, 1859, 1862, 1865 carbon dioxide lasers in, 1853 commercial products using, 1852, 1852 continuous wave lasers in, 1852-53 copper media and, 1851 cost of, 1850, 1854-55, 1865 difference frequency generation laser in, 1853 distributed feedback lasers and, 1853-54 divergence of beam in, 1854, 1859 erbium doped fiber amplifiers in, 1853 Ethernet and, 1851 eye safety with lasers in, 1864-65 fading in, 1862-63 fiber distributed data interface and, 1851 field of view of receivers in, 1859 future of, 1865 GaAs lasers in, 1853 growth of, 1850-51 history and development of, 1850-51 HITRAN database and calculations for, 1855 holographic lens for, 1864 InGaAs lasers and, 1853 lasers and, 1850, 1851, 1852-55, 1853, 1865

lidar lasers and, 1863 light emitting diodes and, 1850, 1852, 1853, 1854, 1865 LOTRAN database and calculations for, 1856-57 Mie scatter in, 1855-57 MODTRAN database and calculations for, 1856-57 modulation tolerance in, 1851 Nd/YAG lasers in, 1853 noise equivalent power in, 1858, 1859, 1859, 1860 noise in, 1857-59 optical detectors and, 1857-59, 1858 optical parametric oscillator in, 1853 power spectral density in, 1862 protocols and, 1851 quantum cascade lasers in, 1853 range equation for, 1859-61 Rayleigh scatter in, 1855-57 receivers for, 1851-52, 1852 reliability of, 1865 resonant scatter in, 1855-57 satellite communications and, 1850 scattering in, 1851, 1855-57, 1856, 1857 scintillations and, 1861-63, 1861 sensitivity in, 1858 signal to noise ratio in, 1858, 1859, 1859, 1860, 1862 SONET and, 1851 telescope design, tracking/alignment, and environment for, 1863-64 thermal noise in, 1858 topologies for, 1851 tradeoffs in design and engineering of, 1865 transmission spectrum for, 1855, 1855 transmitters for, 1851-52, 1852 vertical cavity surface emitting lasers in, 1853 free space propagation equations, 208-209, 209, 2015-16, 2066 free spectral range, 1724, 1787 freespace transmission loss, 2067 frequency and spectrum allocation, 370 frequency assignment problem, cell planning in wireless networks and, 382-383, 382 frequency diversity, 371 microwave and, 2564 mobile radio communications and, 1481 wireless multiuser communications systems and, 1603 frequency divider, frequency synthesizers and, 843-844 frequency division duplex adaptive receivers for spread-spectrum system and, 96 cell planning in wireless networks and, 385-386 frequency division multiple access and, 828 global system for mobile and, 911 IS95 cellular telephone standard and, 347 very high speed DSL and, 2801 frequency division multiple access, 458, 825-830, 2274 acoustic telemetry in, 25, 27 adaptive receivers for spread-spectrum system and, 95–96, 96 admission control and, 120 ALOHA protocols and, 825 alternative implementations for, 828 antenna arrays and, 163 applications for, 829 ATM and, 2907-09 Bluetooth and, 309 carrier sense multiple access and, 349 cellular telephony and, 829 code division multiple access, 829 frequency division duplexing and, 828 frequency division multiple acess frequency hopping spread spectrum and, 828 frequency plan for, 825-826, 826 global system for mobile and, 828, 911-912 interference and, 826-827 intermodulation noise in, 826-827, 827 IS95 cellular telephone standard and, 347, 349 media access control and, 6, 1344 mobile radio communications and, 1481-82, 1482 orthogonal frequency division multiplexing, 828 orthogonal transmultiplexers and, 1880-85

3024 INDEX

frequency division multiple access (continued) packet rate adaptive mobile receivers and, 1886 performance in, 826-827 polyphase sequences and, 1976 powerline communications and, 2003 radio resource management and, 2090, 2091-93 satellite and, 829 satellite communications and, 878-881, 1231-32, 1231, 1231 satellite onboard processing and, 477, 481-482 shallow water acoustic networks and, 2208, 2215 single channel per carrier in, 825 software radio and, 2312-13, 2312 SPADE system in, 825 spatiotemporal signal processing and, 2336 throughput in, 827, 827 time division multiple access and, 828, 829, 2586 underw/in underwater acoustic communications, 44 universal mobile telecommunications service and, 828 wavelength division multiplexing and, 829 wireless local loop and, 2950-51, 2955 wireless multiuser communications systems and, 1602 frequency division multiplexing, 1906 ALOHA protocol and, 130 discrete multitone and, 736–737, **737** flow control and, 1625 multicarrier CDMA and, 1522 optical fiber and, 1709 partial response signals and, 1929 simulation and, 2286 tropospheric scatter communications and, 2693 wavelength division multiplexing and, 2838 frequency domain, in antenna modeling and, 169, 170 frequency domain coding, multicarrier CDMA and, 1524 frequency domain constraints, in optical recording and, 579 frequency domain duplexing, 190 frequency domain equalization, in orthogonal frequency division multiplexing and, 1877 frequency domain equalization, 745 frequency encoded CDMA, 1816–17, **1817**, **1818** frequency hopping, 16, 912, 2092 frequency hopping CDMA, 310, 316, 445, 458, 1344-1345, 1345, 2276 frequency hopped spread spectrum, 309-310, 2216-17, 2396-99 frequency division multiple access and, 828 interference and, 1130-41 wireless communications, wireless LAN and, 1285 frequency independent antennas, 180 frequency modulation, 807-825, 1478, 1825 active antennas and, 50 CDROM and, 1735 community antenna TV and, 519-522, 521 constrained coding techniques for data storage and, 576 continuous phase frequency shift keying and, 593-594 digital audio broadcasting and, 679-680 magnetic storage and, 1327 modems and, 1497 pulse position modulation and, 2033 frequency modulation DCSK, chaotic systems and, 422, 425-427, 426 frequency nonselection (see also flat fading), 784, 1604 frequency offset, in orthogonal frequency division multiplexing and, 1875 frequency range, in powerline communications and, 2000 frequency response community antenna TV and, 517-518, 517 magnetic recording systems and, 2251-52, 2252 orthogonal frequency division multiplexing and, 1874 frequency reuse cellular telephony and, 191, 347, 1479, 1480, 1480

cochannel interference and, 448, 449–454 multibeam phased arrays and, 1514 power control and, 1982

satellite onboard processing and, 479 wireless multiuser communications systems and, 1608 frequency selective channels, in wireless multiuser communications systems and, 1605 frequency selective digital filters, 692-696, 700 frequency selective fading, 2564, 2919 frequency selective switches, WDM, 2840, 2841 frequency shift keying (see also digital phase modulation), 16, 709-719, 2179 acoustic telemetry in, 23, 24 continuous phase frequency shift keying and, 593 high frequency communications and, 947 modems and, 1497, 1498 powerline communications and, 1995 satellite communications and, 1225 shallow water acoustic networks and, 2207 spread spectrum and, 2396-97 Sunda's, 1472 underw/in underwater acoustic communications, 46 underw/in underwater acoustic communications, 40-41 40 frequency slots, flow control and, 1625 frequency synthesizers, 830-865, 831 analog to digital conversion and, 833-835, 834 digital direct, 833-835, 834, 835 digital to analog conversion in, 833-835, 834 double mix divide technique in, 831 fractional N division synthesizers and, 830, 833, 833, 845, 854-862, 860, 861, 862 frequency divider in, 843-844 frequency in, 831-832 frequency pushing in, 836-837 frequency range in, 835-836 harmonic suppression in, 836 hybrid, 830 loop filter in, 845, 846-848 loop gain in, 851-853, 852, 853 oscillator in, 837-843, 838-843 output power in, 836 phase detector in, 844-845 phase locked loop and, 830, 832-833, 833, 845-854, 848 phase noise in, 836, 842–843 sensitivity in, 837 spur suppression in, 858-862 spurious response in, 836 step size in, 836 transient response in, 851-853, 852, 853 tuning and drift in, 837 voltage controlled oscillator and, 830, 836, 843, 860 Fresnel reflection coefficient, 56, 209 Fresnel region, 181-182, 181-182, 182, 214, 1292, 1293, 1438, 1514, 2557–58, **2558** Fresnel ripples, chirp modulation and, 442 fricatives in speech coding/synthesis and, 2360 Friis equation, in indoor propagation models and, 2015 full response continuous phase chirp modulation, 444-446 full response signals, partial response signals and vs., 1928, 1929, **1929** full-duplex acoustic echo cancellation and, 5 fundamental frequency, in speech coding/synthesis and, 2372-73, 2820 fundamental or dominant mode, in waveguides and, 1390 fundamental range, in sampling and, 2107 G.723.1 multimode coder, speech coding/synthesis and, 2354-55, 2354 gain active antennas and, 58 adaptive antenna arrays and, 68 adaptive receivers for spread-spectrum system and, 96, 101, 106 antenna arrays and, 142, 143 antenna modeling and, 169, 170, 185-186, 190, 192-193, 196 antennas for mobile communications and, 196 cable modems and, 327 carrier sense multiple access and, 348-349

chirp modulation and, 443 coding division multiple access and, 458-461 community antenna TV and, 517 dipoles, 1258 diversity and, 729 image compression and, 1063 lasers and, 1778 linear antennas and, 1258 microstrip/microstrip patch antennas and, 1360-1361, 1361, 1380 microwave and, 2570-71 millimeter wave antennas and, 1425 multibeam phased arrays and, 1517 multiple input/multiple output systems and, 1450-1453 optical fiber and, 1842-43 optical fiber systems and, 1842-43 orthogonal frequency division multiplexing and, 1875 packet rate adaptive mobile receivers and, 1888 parabolic and reflector antennas and, 1923-24, 2080-81 path gain and, 1936 satellite onboard processing and, 477 shell mapping and, 2221 space-time coding and, 2324 speech coding/synthesis and, 2347-48 statistical multiplexing and, 2420-32 wireless transceivers, multi-antenna and, 1579, 1583, 1584 gain switched lasers, solitons and, 1771 gain to system noise, 196, 1229, 1927 galactic noise, 949, 2067 Galerkin method, in antenna modeling and, 174, 177 Gallager function, in sequential decoding of convolutional coding and, 2157 Gallagher low density parity check coding and, 658, 659 gallium arsenide, 1742, 1853 Galois fields cyclic coding and, 617, 618, 620 low density parity check coding and, 661 multidimensional coding and, 1538 multidimensional coding and, 1542 optical synchronous CDMA systems and, 1810 ternary sequences and, 2538-47 gamma channels (see also lightpaths), 2098 gap fed loop antenna, 1294, 1295-96 gas lasers, 1777 gatekeepers, IP telephony and, 1174 gateway GPRS support node, 867-876, 2983-84, 2983, 2988 gateways global system for mobile and, 906 IP telephony and, 1174 powerline communications and, 1999 satellite communications and, 881, 882, 1232, 2114 session initiation protocol and, 2198 Gauss elimination, antenna modeling and, 173 Gaussian channels, information theory and, 1114 Gaussian filters, in channel modeling, estimation, tracking, 414 Gaussian frequency shift keying Bluetooth and, 310-311 intelligent transportation systems and, 508 Gaussian memoryless sources in, 641 Gaussian minimum shift keying, 371, 584-593, 718 global system for mobile and, 913 satellite communications and, 1225, 1225 Gaussian random number generation, 2292 Gaussian-Markov noise, in magnetic recording systems and, 2265–66, 2265 general packet radio service, 866-876, 867 admission control and, 126 air interface for, 871-875 architecture for, 866-868 base station controller in, 866-876 base station subsystem in, 866-876 base transceiver station in, 866-876 border gateways in, 867 cell planning in wireless networks and, 369, 383-385, 385 channel coding in, 874-875

general packet radio service (continued) data link layer and, 871-872 dynamic host configuration protocol in, 869 enhanced 866 gateway GPRS support node in, 867-876 global system for mobile and, 866-876, 908 GPRS tunneling protocol in, 867 intelligent transportation systems and, 502, 503, 506–508, **506**, **507** international mobile equipment identity in, 867 Internet and, 866 Internet protocol and, 866 logical channels in, 872-873 microelectromechanical systems and, 2, 1350 mobility portals and, 2191, 2193 packet and circuit switching in, 869 packet data protocol in, 867 protocols for, 870-872, 871 quality of service (QoS), 866, 868-869 radio resource management and, 873-874 satellite communications and, 2117, 2118 security in, 875 services of, 868-869 serving GPRS support node in, 867-876 session, mobility management and routing in, 869-870, 873 short message service and, 866 standards for, 866 subnetwork dependent convergence protocol in subnetwork dependent convergence, 871 support nodes in in, 866-876 time division multiple access and, 2589 universal mobile telecommunications system and, 866 wireless application protocol and, 866 wireless packet data and, 2983-84, 2983, 2988 general source/link algorithm, flow control and, 1628-29 generalization in neural networks and, 1675, 1677-79 generalized Barker sequences, in polyphase sequences and, 1980-1981 generalized chirplike sequence, 1978 generalized Lloyd algorithm, 2128, 2129 generalized minimum distance decoding, 261 generalized minimum shift keying, 1457, 1458 generalized MPLS, 1799 generalized partial response, 4, 1331-33 generalized processor sharing, 1565, 1660 generalized sidelobe canceler, packet rate adaptive mobile receivers and, 1889-90, 1889, 1892-93 generalized simple merging piggybacking, 233 generalized tamed frequency modulation, 585 generating functions, cyclic coding and, 623 generator polynomials, 618, 1610 generic cell rate algorithm, 201, 205-206, 205, 266, 267, 1656, 1659 generic routing encapsulation, 2808 genetic algorithm, 162-163, 163, 2130 geographic diversity, in SONET and, 2495 geographic functions, cell planning in wireless networks and, 374 geographic information system, cell planning in wireless networks and, 372 geographically routed protocols, ad hoc wireless networks and, 2890 geolocation (see also wireless, location in), 2959 geolocation of wireless networks, indoor, 2688-90, 2689 geometric optics parabolic and reflector antennas and, analysis in, 2081-82 path loss and, 1936, 1942 geometric theory of diffraction, 216, 2018 GEOSTAR probe for shallow water acoustic networks and, 2206 geostationary satellite (see also satellite communications), 196, 196, 876-885, 1223, 1224, 1231, 1248, 1250–52, 2113, 2656 GFLOPS processing in software radio and, 2311 Gigabit Ethernet, 1284, 1501, 1508-09, 1510 broadband and, 2655, 2656-58, 2657

dense wavelength division multiplexing in, 1509

erbium doped fiber amplifiers in, 1509 fault tolerance and, 1640-42, 1642 multiwavelength optical network and, 1509 optical fiber and, 1507, 1509, 1721 reliability and, 1640-42, 1642 SONET vs., 1509 standards for, 1509 wavelength division multiplexing in, 1507, 2864 wide area networks and, 1509 Gilbert coding, 1540 global and interleaved constraints, in constrained coding techniques for data storage and, 582 global positioning system antennas and, 169, 198 cdma2000 and, 359 digital audio broadcasting and, 685 Global Positioning System interference and, 1130-41 location in wireless systems and, 2960-61, 2960 millimeter wave propagation and, 1436 satellite communications and, 1224, 1254 spread spectrum and, 2399 wireless sensor networks and, 2994-95 global system for mobile, 96, 308, 369-371, 828, 905-17, 906, 1479, 1480, 2179 air interface for, 908-913 architecture for, 905-907 authentication center in, 906 base station controller in, 905-17 base station subsystem in, 905-17 base transceiver station in, 905-17 blind equalizers and, 296-297 broadband and, 2656 cell planning in wireless networks and, 369-370, 372, 377-383 cellular communications channels and, 397 chann/in channel modeling, estimation, tracking, 409 channel coding in, 910-911 cochannel interference and, 455 connection management in, 915 customized application for mobile network and, 908 enhanced data rate for GSM evolution and, 908 equipment identity register in, 906 frequency division duplexing in, 911 frequency division multiple access and, 911-912 frequency hopping in, 912 frequency planning, 912 gateways in, 906 Gaussian minimum shift keying in, 913 general packet radio service and, 866-876, 908 high speed circuit switched data in, 908 history and development of, 907-908 home location register in, 906 IMT2000 and, 1095-1108 intelligent transportation systems and, 502, 506, 507 international mobile equipment identifier in, 906 international mobile subscriber identity in, 906 logical channels for, 909 media access control and, 1343, 1344 microelectromechanical systems and, 2, 1350 mobile application part in, 906 mobile radio communications and, 1481, 1482 mobile station in, 905-17 mobile telephone ISDN number in, 906 mobility management in, 914-915 mobility portals and, 2192, 2193 modulation in, 913 networking in, 913-916 operation and maintenance, 907 power control in, 913 radio resource management (RRM) and, 914, 2089 regular pulse excitation with long term predictor in, 1304 roaming and handover in, 915-916, 916 routing in, 914-915, 915 satellite communications and, 2116 security in, 916–917 services in, 907–908 session initiation protocol and, 2198 signaling in, 913-916, 913 signaling system 7 and, 906

signaling traffic in, 911 software radio and, 2314 space-time coding and, 2326 speech coding/synthesis and, 909-911, 2356, 2819-20, 2827 spread spectrum and, 2400 standards for, 905 subscriber identity module in, 906 synchronization in, 912-913 time division multiple access and, 911-912, 911, 2589 unequal error protection coding and, 2766-67, 2767 universal mobile telecommunications system and, 907 visitor location register in, 906 wireless application protocol and, 908 wireless IP telephony and, 2932-41 wireless local loop and, 2951-52, 2955 admission control and, 126 antennas, 194 Globalstar in, 196, 1231, 1247, 1250, 1250, 1251, 2112, 2673 go back N ARQ, 226-227, 227, 229-230, 545, 2210 Godard algorithms, in blind equalizers and, 292 Golay coding, 616, 620-621, 885-892 BCH coding, binary, and, 245-246, 247, 251 complementary sequences for, 892-900 deep space telecommunications and, 628-629 Hadamard coding and, 929 low density parity check coding and, 659 peak to average power ratio and, 1950 Golay complementary sequences, 892-900 Golay-Davis-Jedwab coding, 1951 Gold sequences, 428, 900-905 Kasami sequences and, 1219-22 polyphase sequences and, 1976 signature sequence for CDMA and, 2281-82, 2281, 2282 Golomb sequence, polyphase sequences and, 1977 Gordon-Haus effect, 1490, 1767, 1769 GPRS support nodes, 866-876 grace patching, 234 graceful degradation, in media access control, 7, 1345 grade of service, cell planning in wireless networks and, 379-380 gradient method, in antenna arrays and, optimization using, 161 Gram-Schmidt procedure, antenna arrays and, 158 granular noise, waveform coding and, 2835 graph coloring problem, media access control and, 6, 1344 graphs, low density parity check coding and, 1315 gratings, surface acoustic wave filters and, 2446-47, 2446 Gray coding phase shift keying and, 715 photonic analog to digital conversion and, 1961, 1962-1963, 1963 pulse amplitude modulation and, 2027 quadrature amplitude modulation and, 2043, 2044 rate distortion theory and, 2075 serially concatenated coding and, 2173, 2173 trellis coded modulation and, 2625 Gray labeling, bit interleaved coded modulation and, 281, 282, 284-285 Gray mapping pulse amplitude modulation and, 2023, 2023 serially concatenated coding for CPM and, 2187, 2187. 2188 grazing angles, waveguides and, 1416 greedy patching, 234 greedy piggybacking, 233 Green Book, 1736 Green function, in antennas and, 172, 174, 176 Gregorian parabolic and reflector antennas and, 1920-21, 1921, 2083-84, 2083 grid oscillators, in active antennas and, 66, 66 ground reflection point, 209-210 ground wave propagation, 208, 946-958, 2059-60 group blind multiuser detection, 306 group communication and multicasting and, 1529-31, 1529, 1530

group velocity dispersion, 1764, 1765, 1769 guaranteed frame rate, ATM and, 1658 guard bands, in community antenna TV and, 523 guard channels, in admission control and, 124, 124 guard interval, 739-740, 739, 1872 guided scrambling, constrained coding techniques for data storage and, 579 H.261 video codec, 1051-52 H.263 standards, image and video coding, 1052-53 H.323 IP telephony standards, 1173-82, 1175, 2198 H.324 standard, 918-929, 919 Haar transform, image and video coding, 1039 Hadamard coding, 24, 929-935 Hadamard matrices, 898, 1976 Hadamard MFSK, 16 Hadamard transform, 933 Hadamard-Walsh coding, code division multiple access, 2874 half duplex Ethernet, 1504 half power beamwidth, 143, 153, 185, 1358, 1922–23, 1925–26 half wave antennas, 193 half-duplex, acoustic echo cancellation, 5 Hall model of impulsive noise, 2403, 2406-07 Hall's log correlator, for impulsive noise, 2413-14 Hamming coding automatic repeat request and, 225, 229-230 BCH coding, binary, and, 245, 247, 247 cyclic coding and, 617 multidimensional coding and, 1541 product coding and, 2010-11 threshold coding and, 2579-80 underw/in underwater acoustic communications, 43 Hamming distance bit interleaved coded modulation and, 278-279, 281, 282 continuous phase modulation and, 2182 convolutional coding and, 602-604, 603 low density parity check coding and, 1309, 1310 product coding and, 2008 sequential decoding of convolutional coding and, 2142 serially concatenated coding and, 2173 serially concatenated coding for CPM and, 2182 speech coding/synthesis and, 2355 trellis coded modulation and, 2634 Hamming distortion, 640, 2073 handheld device markup language, 2899 handoffs admission control and, 120, 123-126, 124 ATM and, 2912-14 cdma2000 and, 366 cellular telephony and, 1479 global system for mobile and, 915-916, 916 intersatellite handoffs in, 2119 IS95 cellular telephone standard and, 356 radio resource management and, 2093 satellite communications and, 1252, 1254, 2118, 2119-20 wireless multiuser communications systems and, 1602 handover, 2912-14 hands free telephone, 1 handshake protocols shallow water acoustic networks and, 2215-17, 2216 transmission control protocol and, 2607-08 Hankel transforms, impulsive noise, 2404-05 Hansen-Woodyard endfire antenna arrays, 145 hard decision decoding algorithms convolutional coding and, 601-602 low density parity check coding and, 1309, 1312 magnetic recording systems and, 2257 multidimensional coding and, 1541 Reed-Solomon coding for magnetic recording channels and, 475 sequential decoding of convolutional coding and, 2142-45 trellis coding and, 2640 hard disk drives, 1319, 1320-1322 access time of, 1321 areal density of, 1321, 1322

bit error rate in, 1320 capacity of, 1320, 1321 data storage on, 1320-1322, 1322 data transfer rates in, 1322 extended partial response in, 1328-1331, 1329, 1331, 1332 form factors in, 1320, 1322 head space in, 1320 latency of, 1321 linear density of, 1321 partial response maximum likelihood in, 1321, 1328, 1328, 1330-1331 RAMAC systems in, 1320, 1321 read process in, 1320 redundant array of independent disks and, 1322 seek time in, 1320-1321 synchronization in, 1320 track density of, 1321 tracks on, 1320-1321 trends in, 1320-1321 volumetric density of, 1321-1322 write process in, 1320 zone bit recording in, 1321 hard handoff, in cdma2000, 366 hard limiters, in optical synchronous CDMA systems, interference cancellation, 1821-23, 1821, 1822, 1823 hardware description language, 2285 harmonic broadcasting, 236 harmonic suppression, in frequency synthesizers, 836 hash functions, 221-222, 612-613, 1152 Hasse-Weil theorem, in cryptography, 610 Have Quick (see also software radio), 2310-12 head end (HE) cable modems, 324, 512, 513 head noise, in digital magnetic recording channel, 1325 head space, in hard disk drives, 1320 header error control, 200, 201, 312, 550, 2977-78 headers, ATM, 200, 550 heavy tailed on/off models, in traffic modeling, 1669 hectometric (see medium frequency) helical antennas, 180, 193-194, 193, 935-946, 936-945 Helmholtz (scalar wave) equation, 171, 1394 hemispherical conformal antenna arrays, 152-153 hertz, 1423 Hertz active antenna, 48-68 Hertz, Heinrich, 48, 179, 208, 370, 677, 1477, 2585 heterodyne receivers, in optical transceivers, 1835 heterostructure lasers, 1777, 1777 heuristic algorithms in quantization, 2128 heuristic function, in sequential decoding of convolutional coding, 2145-46 HF data link, 947 hidden Markov model, 958-966, 959 applications for, 964-965 automatic speech recognition and, 2373-80, 2385 Baum-Welch algorithm in, 961-962 blind equalizers and, 290-291 chann/in channel modeling, estimation, tracking, 406 Markov chain and, 963 maximum likelihood and, 961-962 maximum likelihood estimation and, 1341 maximum mutual information in, 962 Viterbi algorithm and, 961, 2818, 2818 hidden node problem, 1286-87, 1286 hidden terminal problem ad hoc wireless networks and, 2885, 2885 carrier sense multiple access and, 345-346, 345 media access control and, 1343, 1347 hierarchical forwarding, in multiprotocol label switching, hierarchical routing, 1566, 2890 high- and low-latitude curve and skywaves, 2061 high data rate packet transmission, cdma2000, 368 high definition TV, 966-979 BISDN and, 263 chann/in channel modeling, estimation, tracking, 402 digital versatile disc and, 1738 tapped delay line equalizers and, 1689 high density bipolar 3, 1934 high frequency, 946–958, 2059–60 high latitude anomalies, 2065–66

high level data link control, 546-547, 546 asychronous balanced mode in, 546 asynchronous response mode in, 546 bit stuffing and, 547 frame check sequence in, 547 information frames in, 546 normal response mode in, 546 supervisory frames in, 546 unnumbered frames in, 546 high order sequence criteria, 291-292, 291 high rate punctured convolutional coding, 979-993 High Sierra File format, 1736 high speed circuit switched data, 383-385, 908 high speed DSL, 317 high speed photodetectors for optical communications, 993-1006 higher data rate, in admission control, 126 high-gain antennas, 169 highly elliptical orbit satellite, 1249 highpass filters, waveguides as, 1390 hijacking, 1646, 2810 Hilbert transform, amplitude modulation, 135 Hill plots, interference, 1123-24 HiperAccess group, broadband wireless access, 319-320 HiperLAN, 308, 2682, 2683, 2684, 2941, 2945 ATM and, 2909 broadband wireless access and, 320-321 media access control and, 1348 wireless LANs and, 2681, 2682 HiperMAN, in broadband wireless access, 320 histogram algorithm (HA), in quadrature amplitude modulation, 2056 histogram evaluation, in signal quality monitoring, 2270-71, 2271 history, in constrained coding techniques for data storage, 578 HITRAN database and calculations for free space optics, 1855 Hobbs coding, 1540 holding time, traffic engineering, 485 hole punchers, impulsive noise, 2416 holographic concave gratings, 1755, 1755 holographic data storage system, 1740 holographic memory/optical storage, 1740, 1740, 2132-35, **2133**, **2134**, **2135** bit error rate in, 2138 charge coupled devices in, 2134-35 decryption system for, 2137-38, 2137 double random phase encryption in, 2132-33 encryption, cryptography and, 2132 experimental setup for, 2134-35, 2134 free space optics and, 1864 holographic memory and, 2132-35, 2133, 2134, 2135 lasers in, 2134 numerical evaluations of, 2138 plane waves and, 2133 random phase mas in, 2133 receiver for, 2136-37, 2137 sampling and, 2138 transmitter for, 2135-36, 2135 home area network, 2685-88, 2687 home computing home area network, 2685-88, 2687 home location register, 906, 2987 home networking, wireless, 2684-88, 2685 home phone network of America, very high speed DSL, 2790 Home RF, 1289, 2683, 2684 homodyne receivers, in optical transceivers, 1835 hop, in media access control, 6, 1344 hop by hop protocols, 116, 541 hop selection, Bluetooth, 312-313 Hopfield neural networks, 1677 hopping, high frequency communications, 949 horn antennas, 142, 179, 180, 184, 187, 1006-17, 1006, 1392, **1392**, 1425–28, **1427**, **1425** host identifier, 548 Hotelling transform for waveform coding, 2837 hubbed architectures, Ethernet, 1505, 1505 Huffman coding, 1017-24 compression and, 634-635, 637

Huffman coding (continued) image and video coding and, 1031-32 scalar quantization and, 2124 human made noise, 949, 2067 Huygen's principle, 183-184, 214 hybrid fiber coax systems, 512, 518-522, 518 hybrid IntServ-DiffServ, 271 hybrid optical networks, medium access control, 1559 hyperplanes finite geometry coding and, 802 low density parity check coding and, 661 hypertext markup language, 2900 hypertext transfer protocol, 2199, 2203, 2604, 2899 IATSAMTR scheduling in medium access control, 1558 idle handoff, cdma2000, 366 image and video coding, 1025-62 additive white Gaussian noise, 1034 advanced video coding in, 1054-55, 1055 analog to digital conversion in, 1026-27 arithmetic coding in, 1032 bit allocation and rate control in, 1044-46, 1045 block coding in, 1038-39 color space in, 1026 color subsampling in, 1026 compression in, 1030 compression ratios in, 1028-29 context formation in, 1047 differential PCM and, 1037-38 discrete cosine transform in, 1039 discrete waveform transform in, 1040 entropy coding in, 1031-33 error detection and correction in, 1030-31 evaluating schemes for, 1028-29 fractal compression in, 1044 future research in, 1055-57 generic model for, 1029-31, 1029 H.261 video codec in, 1051-52 H.263 standards in, 1052-53 Haar transform in, 1039 Huffman coding in, 1031-32 imaging in, 1025-26, 1026 interlaced scanning in, 1027 JBIG standards for, 1049 JPEG compression and, 1029, 1049-50, 1050, 1211-18 Karhunen-Loeve transform in, 1039 lapped orthogonal transforms in, 1039 Lempel-Ziv coding in, 1032-33 linear transformations in, 1038-42 mapping in, 1030 Markov sources in, 1033 motion estimation and compensation in, 1042-44, 1042 MPEG compression in, 1029, 1052, 1053-55, 1054 object-based coding in, 1057, 1057 post processing in, 1030–31, 1047–48 pre processing in, 1047–48 predictive coding in, 1033-34, 1034, 1037-38, 1037 quantization in, 1026-27, 1030, 1035 redundancy and irrelevancy in, 1027-28 run length coding in, 1030, 1046 scalable coding in, 1056-57 scalar quantization in, 1035 shape adaptive transforms in, 1041-42 signal models for, 1034-35 standards for, 1048-55 subband decomposition in, 1039-41, 1041 symbol formation in, 1046-1047 transcoding in, 1057 uncompressed digital video in, 1027 variable length coding in, 1030 vector quantization in, 1030, 1035-37, 1036 vector transformations and, 1041 video scanning and frame rate in, 1027 visual texture coding in, 1050 zero tree coding in, 1046-47 zigzag scanning in, 1046 image compression, 1062-73 image processing, 1073-79 image sampling and reconstruction, 1079-94, 1081-92

image source, in acoustic echo cancellation, 3 image transmission, video, unequal error protection coding, 2764–65, **2765** iMode, 2193 impedance, impedance matching active antennas and, 49, 50, 50, 56 antenna arrays and, 160 antennas and, 169, 177, 177, 180, 184, 186 community antenna TV and, 524 dipoles, 1258 impedance, impedance matching linear antennas and, 1258 loop antennas and, wave impedance in, 1293-95, 1292 microstrip/microstrip patch antennas and, 1358, 1359, 1360, 1362, 1363, 1383 powerline communications and, 2000 television and FM broadcasting antennas, 2517-36 waveguides and, 1395, 1395, 1398-99, 1399, 1401-03, 1403 waveguides and, 1398-99, 1399 impersonation attack, 219, 222 importance density function, in speech coding/synthesis, 2365-66, 2366 impulse response, acoustic echo cancellation, in LEMS, 2, 2, 3, 1689, 1689 impulsive noise, 2402-2420 IMT2000, 358, 392, 1094-1108, 2873-74 admission control and, 126 advanced mobile phone service and, 1095-1108 architecture for, 1101-06 cdma2000 and, 1096-1108 cell planning in wireless networks and, 369, 386 cellular communications channels and, 397 code division multiple access, 1095-1108 context and evolutional paths of, 1097-99 core network for, 1102-04 customized applications of mobile network enhanced logic and, 1101-08 digital enhanced cordless telephony and, 1096-1108 enhanced data for GSM evolution and, 1096-1108 future of, 1106-07 global system for mobile and, 1095-1108 harmonization efforts in, 1097 history and development of, 1097-99 IP networks and, 1103 license assignment and economic implications of, 1106 migration paths for, 1098-99, 1100 mobile IP and, 1103 open service architecture for, 1100-01 Parlay and, 1100 quality of service, 1099-1101, 1103 radio access network for, 1101-02, 1102 radio interface for, 1095-96 satellite communications and, 2116 standards for, 1095-97 synchronous CDMA and, 1096 terminals and services for, 1099-1101 terminals, 1101, 1101 time division multiple access and, 1095-1108 universal mobile telecommunications system and, 1096-1108 virtual home environment for, 1101 wideband CDMA and, 1096, 1104-05 wireless application protocol and, 1100 wireless IP telephony and, 2932-41 in band signaling, transport protocols for optical networks, 2618 inband interference, predistortion/compensation in RF power amplifiers, 530 incliometers, 20 incoming label map table, in multiprotocol label switching, 1591, 1592, 1594 incremental redundancy principle, in automatic repeat request, 231 independent BSS, 1285 index calculus method, in cryptography, 610 index of dispersion interval, statistical multiplexing, 2424 refraction (see refractive index)

index, antenna (see antenna index) indexing, shell mapping, 2223 individually optimal detector, 99 individuals, in quantization, 2130 indoor propagation models, 2012-21 advanced ray optical model and database preprocessing in, 2019-2020, 2020 antennas and, 2015 categories of environments in, 2014 database for buildings and their characteristics in, 2014-25, 2014 deterministic models in, 2018-20 diffraction and, 2013, 2018 dominant path concept in, 2020, 2021 empirical narrowband model in, 2015-17, 2015 empirical wideband models in, 2017-18 fading and, 2013, 2013 free space model in, 2015-16 Friis equation in, 2015 geometric theory of diffraction in, 2018 mateial properties and, 2013-14 mobile indoor radio channel in, 2013 Motley-Keenan model in, 2016, 2016 multipath in, 2013, 2013, 2018 multiwall model in, 2016-17, 2017 one slope model in, 2016, 2016 path finding in, 2019, 2019 path loss in, 2015 penetration in, 2013, 2018 planning tools for, 2020 power delay profiles in, 2017 ray launching, ray tracing in, 2019, 2019 reflection and, 2013, 2018 scattering and, 2013, 2018-19 transmitter and receiver location in, 2013 ultrawideband radio and, 2758-59 universal theory of diffraction in, 2018 visibility relations in, 2020, 2020 wide area networks and, 2012 indoor wireless networks, 734-735, 2677-92 induced local fields, in neural networks, 1676 inductance, 55, 2000 induction zone, loop antennas, 1292-93, 1293 inductors, in microelectromechanical systems, 1352-1353, 1353 industrial scientific medical band, 309, 316, 2391 infinite antenna arrays, 165-166 infinite impulse response filters, 694-696, 695, 697-698 simulation and, 2288, 2288 speech coding/synthesis and, 2343, 2346 acoustic echo cancellation and, 3 in channel modeling, estimation, tracking, 415 infinite length symbol spaced equalizers, 1690-92 infinitesimal generator model, for traffic engineering, 488-489 infinity norm method, in peak to average power ratio, 1947-48 information frames, in high level data link control, 546 information rate, for satellite communications, 1229 information theory, 1109-15 InfoSec software radio, 2307, 2308, 2313 Infrared Data Association, 2041 infrared optical fiber, 434 infrared transmission, 2925-31, 2925 Bluetooth and vs., 307 intelligent transportation systems and, 504-505 pulse position modulation and, 2041 InGaAs lasers, 1742, 1853 ingress edge routers, 1802, 1803 ingress noise, in community antenna TV, 524 ingress routers, in differentiated services, 668 inline optical amplifiers, 1710 Inmarsat, 196, 198, 876, 1224, 1227, 2112 inner coding, in serially concatenated coding, 2164 inner cyclic block coding, in underwater acoustic communications, 43 InP 1742 inphase-quadrature signal, in minimum shift keying, 1457, 1459-61, 1460, 1461 1463, 1472 input signal matrix, acoustic echo cancellation, 7 inquiry mode, in Bluetooth, 311-312

insertion loss microelectromechanical systems and, 6, 1354 optical couplers and, 1699 optical cross connects/switches and, 1784 optical fiber and, 1843 optical filters and, 1732 optical modulators and, 1743 instant messaging, session initiation protocol, 2203 instantaneous frequency pulse, 1458 instantaneous narrowband interference, 1130-41 integral approach to antenna modeling, 170, 172-176 integrated circuits fabrication techniques for, 3, 1351 microelectromechanical systems and and, 1350-1356 integrated digital networks, 1567 integrated services (IntServ) admission control and, 114-115, 115 flow control, traffic management and, 1654, 1657, 1657 hybrid IntServ-DiffServ in, 271 IP networks and, 269-270 IP telephony and, 1180 mobility portals and, 2195 multiprotocol label switching and, 1597 integrated services broadcasting system, 680 integrated services digital broadcasting, 2549, 2551-52 integrated services digital network, 263, 1567 H.324 standard for, 918-929, 919 IP telephony and, 1177 modems and, 1495 multimedia networks and, 1563 statistical multiplexing and, 2424 very high speed DSL and, 2780 virtual private networks and, 2808 integrated services LAN, 1641 integrity of data, 1151, 1152, 1648, 1649 intelligent antenna arrays, 163 intelligent networks, 719-29, 722, 726 intelligent noise, 218 intelligent transportation systems, 502-512 Intelligent Vehicle Initiative, 503 Intelsat, 876-885, 1392, 1392 intensity modulation/direct detection, 1809, 1814 interactive voice response systems, 2384 interarrival times, in traffic engineering, 489 intercell interference, in polyphase sequences, 1975 interception attacks, 1151 interdigital transducers, in surface acoustic wave filters, 2447-48, **2448** interdomain multicast routing protocols, 1535-37 interexchange carrier for IP telephony, 1177 interface message processors, 267-268 interference, 208, 218, 1115-21 adaptive antenna arrays and, 68-71 adaptive equalizers and, 79 adaptive receivers for spread-spectrum system and, advanced mobile phone system and, 1130-41 Bluetooth and, 309 broadband wireless access and, 318-319 cell planning in wireless networks and, 377-380 cochannel (see cochannel interference in digital cellular TDMA networks), 448 code division multiple access, 458, 1116, 1119, 1130-41 digital audio broadcasting and, 677 Eigen algorithm for, 1116-19 expectation maximization algorithm and, 772-773 frequency division multiple access and, 826-827 frequency hopping spread spectrum, 1130-41 global positioning system and, 1130-41 Hill plots in, 1123-24 instantaneous narrowband interference, 1130-41 interferer multiplication in, 108 local multipoint distribution service and, 318-319, 1268 microwave and, 2566-67 modeling of, 1121-30 multicarrier CDMA and, 1527 multiple input/multiple output systems and, 1119, 1450, 1452, 1452

multitone, 1130-41 narrowband interference, 1130-41 optical communications systems and, 1484 optical fiber and, 1484 optical filters and, 1723, 1756-57 optical multiplexing and demultiplexing and, 1749 orthogonal frequency division multiplexing and, 1874, 1876 packet rate adaptive mobile receivers and, 1886 partial band, 1130-41 predistortion/compensation in RF power amplifiers and, 530 QQ estimator in, 1124 satellite communications and, 1251 satellite onboard processing and, uplink, 477-478, 478 software radio and, 2306 space-time coding and, 2324 spread spectrum and, 1130-41, 2393-94 time division multiple access and, 1130-41 underw/in underwater acoustic communications, cancellation in, 44 wireless LANs and, 2678 wireless multiuser communications systems and, 1604 wireless networks and, 121 wireless packet data and, 2982 wireless systems and, 1115-21, 1121-30 wireless transceivers, multi-antenna and, 1579 interference cancellation coding division multiple access and, 1817-23, 1819-23 multibeam phased arrays and, 1519, 1520-21 optical synchronous CDMA systems and, 1817-23, **1819**–23 wireless multiuser communications systems and, subtractive and successive, 1617 interference fading, 2065 interference filters, 1723-27, 1723, 1726, 1749 interference function, in power control, 1985 interference rejection, in IS95 cellular telephone standard, 350 interferer multiplication, in adaptive receivers for spread-spectrum system, 108 interferometers Mach-Zehnder (see Mach-Zehnder interferometer) Michelson (see Michelson interferometers) Sagnac (see Sagnac interferometers) interframe spacing, 9, 1347 interior gateway protocol, 269, 1913, 2462 interior router protocols, 549 interlaced scanning, in image and video coding, 1027 interleavers/interleaving, 1141-51, 1142-49 bit interleaved coded modulation and, 276-286, 276 cdma2000 and, 359 CDROM and, 1735 community antenna TV and, 526 concatenated convolutional coding and, 557-558 constrained coding techniques for data storage and, 582 diversity and, 733 high frequency communications and, 954–955 IS95 cellular telephone standard and, 350, 352, 354 magnetic storage and, 2, 1330 peak to average power ratio and, reduction of, 1949-50 Reed-Solomon coding for magnetic recording channels and, vs. noninterleaving in, 472 serially concatenated coding and, 2164, 2165-66, 2165 serially concatenated coding for CPM and, 2183 turbo coding and, 1141–51, **1142–1149** turbo trellis coded modulation and, 2740-42, 2741 intermediate circular orbit systems, 1224 intermediate frequency (IF), 330-334, 1478 intermediate system to intermediate system, 269, 1658 intermediate systems, 549 intermittent failures, 1631 intermodulation, community antenna TV, 512, 514

intermodulation distortion, 191, 328, 330, 826-827, 827, 2697.2698

internal network to network interface (INNI), 1799 International Maritime Satellite System (see INMARSAT), international mobile equipment identifier, 867, 906 international mobile subscriber identity, 906 International Mobile Telecommunications 2000 (see IMT2000 International Telecommunications Satellite Organization (see INTELSAT) International Telecommunications Union standards (see standards), 4 Internet, 115-116, 1909, 2653 access control in, 1650-51 admission control and, 114 automatic repeat request and, 224-231 general packet radio service and, 866 IP networks and, 267, 268 microelectromechanical systems and, 1, 1349 mobility portals and, 2190-91 multicasting and, 1531-32, 1531 multimedia networks and, 1567-69 optical, 2461-72 packet switched networks and, 1912-13 satellite communications and, 2113-15, 2114, 2115, 2120-21 security in, 1151–57, 1650–52 wireless IP telephony and, 2931–41 internet control message protocol, 1646, 2988 internet gateway protocols, 1658 Internet integrated services architecture, 1567 Internet key exchange protocol, 1153-54, 1651, 2812-14, 2813 Internet protocol (see also IP networks), 268, 541, 542-543 broadband and, 2662 flow control, traffic management and, 1653 general packet radio service and, 866 IP telephony and, 1172-82, 1173 microelectromechanical systems and, 1350 multiprotocol label switching and, 1590-1601 packet switched networks and, 1911 satellite communications and, 1253 virtual private networks and, 2809-14 wavelength division multiplexing and, 2864 wireless IP telephony and, 2931–41 Internet relay chat, 2192 Internet Research Task Force, 1647 Internet Security Association and Key Management Protocol, 2813-14 Internet service providers, 2462 IP telephony and, 1177 modems and, 1498-99 satellite communications and, 2115 virtual private networks and, 2808 internetwork layer, TCP/IP model, 541 internetworking, wavelength division multiplexing, 654 internetworking protocols, 547-550 internetworking units, 2116, 2117 interrupt coalescing, in transport protocols for optical networks, 2620 interrupted Bernoulli process, 117 interrupted fluid process, 117 interrupted Poisson Process, 117 interruptive attacks, 1151 intersatellite handoffs, 2119 intersatellite links, 1224, 1252, 2113 intersymbol interference, 208, 1157-62, 1158-61 acous/in acoustic modems for underwater communications, 15 adaptive equalizers and, 79-81, 87 adaptive receivers for spread-spectrum system and, 103 blind equalizers and, 286, 291 blind multiuser detection and, 303 cable modems and, 327, 328 chann/in channel modeling, estimation, tracking, 398, 410, 411, 417 chirp modulation and, 443, 446 code division multiple access, 2278, 2283 digital magnetic recording channel and, 1325-26 magnetic recording systems and, 2251-52, 2251

intersymbol interference (continued) magnetic storage and, 1327, 1329, 1330-1331 minimum shift keying and, 1466-67 mobile radio communications and, 1481 multiple input/multiple output systems and, 1455 optical fiber and, 1970 orthogonal frequency division multiplexing and, 1867 packet rate adaptive mobile receivers and, 1887, 1899 partial response signals and, 1928–35 quadrature amplitude modulation and, 2045 shallow water acoustic networks and, 2207 signature sequence for CDMA and, 2278, 2283 space-time coding and, 2327 spatiotemporal signal processing and, 2333, 2336 synchronization and, 2474-85 tapped delay line equalizers and, 1688 terrestrial digital TV and, 2548 tropospheric scatter communications and, 2699 turbo equalization and, 2716-27 in underwater acoustic communications, 38, 44 Viterbi algorithm and, 2817-18, 2817 wireless multiuser communications systems and, 1612, 1616 intertrack interference (see crosstalk) interuser interference, in coding division multiple access, 461 interval usage coding, in cable modems, 334 interval, sampling, 2106, 2107 INTRA wireless MPEG 4 videocommunications, 2978 intrachannel cross phase modulation, 1769-70 intrachannel four wave mixing, 1769-70 intradomain multicast routing, 1533-35 intranets, 1163-72, 1165, 2807 intrusion detection and response, 1651-52 intrusion detection exchange format, 1651 intrusion detection systems, 1651-52 invariance, in trellis coded modulation, 2632-33 inverse bending (earth bulge), microwave, 2559 inverse discrete Fourier transform orthogonal frequency division multiplexing and, 1871 - 72peak to average power ratio and, 1947, 1950-51, 1951 inverse fast Fourier transform, 532 inverse Fourier transform, 2107 inverse scattering transform, 1766 inverse transform method, 2292, 2292 IONCAP software for radiowave propagation, 2066 ionosphere, 208, 2067 ionospheric propagation, 758-780, 2059, 2060, 2065 ionospheric scintillation, 196, 197 IP addressing, 269, 541, 548-549 mobility portals and, 2194 multicasting and, 1531 packet switched networks and, 1912-13, 1912 satellite communications and, 2117 session initiation protocol and, 2197 wireless packet data and, 2988 IP datagrams, 542, **543** IP host to IP host, 1177–78 IP in IP encapsulation, 2808 IP networks ALOHA protocols and, 268 ARPANET as, 267-268 autonomous systems and, 269 best effort forwarding in, 269 BISDN and, 267-271 border gateway protocol and, 269 broadband and, 2661-64 cdma2000 and, 359 cell planning in wireless networks and, 392 classless interdomain routing and, 269 connectionless nature of, 269 datagrams in, 269 differentiated services in, 270-271 exterior gateway protocols for, 269 forwarding in, 269, 1591 history of, 267–269 hybrid IntServ-DiffServ in, 271 IMT2000 and, 1103 integrated services and, 269-270

intelligent transportation systems and, 507 interface message processors and, 267-268 interior gateway protocols for, 269 intermediate system to intermediate system and, 269 Internet and, 267, 268 Internet protocol and, 268 IP addressing and, 269 IP telephony and, 1172–82, **1173** layer 3 protocol and, 269 metropolitan area exchanges and, 268 multicasting and, 1531 multiprotocol label switching and, 271, 1590–1601 network access points and, 268 network IDs in, 269 open shortest path first and, 269 optical cross connects/switches and, 1798 packet switched networks and, 1910, 1912 packets in, 267 peer networks, bilateral public peering and, 268 point of presence and, 268 quality of service and, 269-271 resource reservation protocol in, 270 routing tables and, 269 satellite communications and, 268, 2111-22 satellite onboard processing and, 482-483 TCP/IP and, 267, 268 transmission control protocol and, 268 tunneling in, 273 virtual private networks and, 273, 2808 Voice over IP, 274 wavelength division multiplexing and, 2845 wireless IP telephony and, 2931-41 IP over ATM, 1154 IP over WDM, 1798, 1799 IP telephony, 1172–82, **1173** IPSec, 273, 1153–54, 1651 IP telephony and, 1180-81 virtual private networks and, 2810-14, 2810 iPSTAR, 2112 IPv6, 2194, 2197, 2663 IrDA, 2682, 2925, 2929 Iridium, 196, 484, 1247, 1250, 1250, 1251, 1253, 1519, 2112 irreducible polynomials, in BCH coding, binary, 241-242 IS136, 370 cell planning in wireless networks and, 383 in channel modeling, estimation, tracking, 409 wireless local loop and, 2952 IS95 cellular telephone standard, 7, 347-358, 370, 1345, 2391, 2400 acquistion and tracking in, 354 advanced mobile phone system, 347 bandwidth in, 350 base station diversity in, 355-356 binary phase shift keying and, 354 cdma2000 and, 357, 367 Cellular Telecommunications Industry Association and, 347 coding division multiple access and, 347-348 convolutional coding and, 353 cyclic redundancy check in, 353 diversity features of, 355-356 electronic serial numbers and, 355 error control in, 350, 354 evolution of, to third-generation, 356-357 finite impulse response in, 350 forward and reverse link channels in, 349-357, 349 forward fundamental/supplemental coding channels in, 356 frequency division duplexing and, 347 frequency division multiple access and, 347, 349 handoff in, 356 interference rejection and, 350 interleaving in, 350, 352, 354 location in wireless systems and, 2965 Mobile Station Base Station Compatibility Standard for Dual Mode Wideband...Cellular, 347 modulation in, 350, **351**, **353**, 354 multipath diversity in, 355-356 multiple access in, 354 multiplexing in, 350, 352

offset quadrature phase shift keying in, 354 paging channels in, 349, 352 Personal Communications Services and, 350 pilot channels in, 349 pseudorandom noise coding in, 349, 350-351, 354 pulse shaping in, 350, 354 quadrature phase shift keying in, 350 quadrature spreading in, 354 RAKE receivers and, 356 reverse link in, 353-355, 355 spread spectrum and, 347 superfinger demodulator for, 357 synchronization channels in, 349, 350 traffic channels in, 349 voice coding in, 350, 354 Walsh functions and, 350, 354, 356-357 wireless IP telephony and, 2932-41 wireless local loop and, 2952 isolation, in multimedia networks, 1564 isotropic radiator antennas, 185 ISUP, 2197, 2198 iteration in antenna arrays, modified patterns by, 156-157, 157 iterative algorithms, in serially serially concatenated coding, 2164-78 iterative coding/decoding (see also concatenated convolutional coding and iterative decoding) 556-570, 565, 567, 1182-96 Bahl-Cocke-Jelinek-Raviv decoding and, 561-564, 564 bit interleaved coded modulation and, 284 compression and, 635-636 continuous phase modulation and, 2180-81 direct sequence CDMA and, 1196 finite geometry coding and, 805-806 log a posteriori probability in, 561 log likelihood ratio algorithm in, 561 low density parity check coding and, 1309, 1311-12, 1316 maximum a priori algorithm in, 561 multidimensional coding and, 1538, 1543, 1544 parallel concatenated convolutional coding and, 564-567, 565 product coding and, 2010-11 satellite communications and, 1229-30, 1230 serial concatenated convolutional coding and, 567-569, 567 serially concatenated coding for CPM and, 2180-81 soft in/soft out decoders and, 560, 564-567 soft output decoding algorithms and, 2295 trellis coding and, 560, 2640, 2650–51 turbo decoders as, 560 turbo product coding and, 2727-37 Viterbi algorithm and, 560 iterative detection algorithms, 1182-1210 iterative least squares with enumeration, blind equalizers, 292 iterative methods in quantization, 2128 Ito-Stratonovich chaotic integrations, 422 ITS irregular terrain model, in radiowave propagation, 216 JAIN, 727, 728, 2203 Jakes model, in channel modeling, estimation, tracking, 410, 411 jamming, 74, 74, 1281, 1283 JBIG standards, image and video coding, 1049 Jini mobility portals, 2194 iitter ad hoc wireless networks and, 2888 broadband and, 2655 cable modems and, 328, 329 constrained coding techniques for data storage and, 573 differentiated service and, 674, 674 flow control, traffic management and, 1660 IP telephony and, 1172-82, 1173 multimedia networks and, 1562 optical fiber and, 1767, 1769 optical signal regeneration and, 1759 solitons and, 1767, 1769

3030 INDEX

Johnson noise, in free space optics, 1858 joint position and amplitude search, 1306 joint source and channel optimization, 1571 joint tactical radio system, 2306, 2311 jointly optimal detector, in adaptive receivers for spreadspectrum system, 98-99 Jones matrix, 436, 1492 JPEG compression, 1076, 1211-18 image and video coding and, 1029, 1049-50, 1050 image compression and, 1063, 1070-72, 1072 magnetic storage and, 1319 JPEG2000 (see JPEG compression) jukeboxes, optical, 1733 Ka band, 877, 1251, 2113, 2665-66 Kalman model, in channel modeling, estimation, tracking, 411-416, **411** Karhunen-Loeve transform compression and, 648 image and video coding and, 1039 image compression and, 1065 transform coding and, 2599–2600, **2599** waveform coding and, 2837 Karn's algorithm, in transmission control protocol, 2612 Kasami sequences, 1219-22 BCH coding, binary, and, 247, 251 Gold sequences and, 904 signature sequence for CDMA and, 2282 Kathryn orthogonal frequency division multiplexing, 1867 keep alive session initiation protocol, 2202 Keller cone for diffraction, 1942 Kerberos, 1155 Kerr effect, 1739, 1765 key distribution center, 1152 key exchange cryptography, 610–611 keyed hash MAC cryptography, 613 kilometric wave (see low frequency) Kineplex orthogonal frequency division multiplexing, 1867 Kirchhoff's current law, in antennas, 175, 1437 Kirke method and groundwave propagation, 2060 Kirkman system low density parity check coding, 659, 664-65 KMB algorithm for multicasting, 1533 k-means clustering algorithm in adaptive receivers for spread-spectrum system, 103 knife edge diffraction, 214-215 Kraft's inequality in compression, 633, 635 Kraft-McMillan inequality in constrained coding techniques for data storage, 577-578 Kronecker product, 2009-2010, 2038 Kruskal's algorithm for multicasting, 1532 KTMTR scheduling, medium access control, 1558 Ku band, 877, 1251, 2113 kylstron antennas, 179 L band satellite communications, 196, 1251 L stage vector quantization, 2127 L step orthogonalization, in threshold coding, 2580 label distribution protocol, 116, 1568, 1591, 1593, 1595-96 label edge routers, 1591 label switched path, 116, 271, 1590-93, 1593, 1798-99, 1802 label switched routers, 116, 271, 1591, 1594, 1798 labels, 549, 1591 labels, security, 1649 lambda switch capable interfaces, 1799 Lambdanet, 1720 LAN emulation, 1719 land mobile satellite communications, 1223-34 Advanced Communications Technology Satellite in, 1227, 1228 ALOHA protocols in, 1232 antenna direction and, 1223, 1228–29 automatic repeat request in, 1223, 1229-31, 1230, 1231 binary frequency shift keying in, 1225, 1225 binary phase shift keying in, 1225, 1225, 1230 bit error rate in, 1224, 1225, 1227, 1230

block coding in, 1229-30, 1230 channel characteristics in, 1224-29 code division multiple access in, 1231-32, 1231 continuous phase modulation in, 1225 convolutional coding in, 1229-30, 1230 differential phase shift keying in, 1225, 1225 diversity techniques in, 1230-31 downlinks in, 1223, 1223 error detection and correction in, 1223, 1229-31, 1230, 1231 fading in, 1223, 1226-29, 1226, 1227 figure of merit in, 1229 file transfer time in, 1233, 1233 forward error control in, 1223, 1229-31, 1230, 1231 forward/reverse path in, 1223, 1223 frequencies for, 1223-24, 1224 frequency division multiple access in, 1231-32, 1231 frequency shift keying in, 1225 gain to system noise in, 1229 gateways in, 1232 Gaussian minimum shift keying in, 1225, 1225 geostationary satellite in, 1223, 1224, 1231, 1232 global positioning system and, 1224 information vs. coding rate in, 1229 intermediate circular orbit systems in, 1224 International Maritime Satellite System and, 1224, 1227 internetworking in, 1232-33 intersatellite links in, 1224 iterative coding in, 1229-30, 1230 Lincoln Laboratory Link Layer protocol in, 1233 link budget for, 1229 low earth orbit satellite in, 1223, 1224, 1231, 1247-56 maximal ratio combinining in, 1230 medium earth orbit satellite in, 1223, 1224, 1231, 1233 modulation in, 1225, 1225 multipath fading in, 1226–27, **1226** multiple access in, 1231–32, **1231** network aspects in, 1232-33 noise in, random, 1224-25 path loss in, 1223, 1225-26, 1225 phase shift keying in, 1225 propagation in, 1223 Rayleigh channels, Rayleigh fading in, 1226-27, 1226, 1227 reservation protocols in, 1232 Rice factor in, 1226-27, 1226, 1227 satellite diversity in, 1223, 1231 satellite transport protocol in, 1233 shadowing in, 1223, 1227-28 signal to noise ratio in, 1224, 1230 slant range in, 1225–26, **1226** space communications protocol standards in, 1233 spatial diversity in, 1230-31 TCP/IP and, 1232-33 time division multiple access in, 1231-32, 1231 transmission control protocol and, 1233 turbo coding and, 1229-30, 1230 uplinks in, 1223, 1223 wireless IP suite enhancer and, 1233 wireless transmission control protocol in, 1233 land use/land clutter, 2561 landmark ad hoc routing ad hoc wireless networks, 2890 lands, CD, 1736 language models, in automatic speech recognition, 2376–77, 2385–77, 2388–89 LAP-B, 546 LAP-D, 546 lapped orthogonal transforms, 1039 laser chirp, optical fiber systems, 1844 laser communications chaotic systems and, 428-431 chirp modulation and, 447 laser intensity noise, optical fiber systems, 1843 laser phase noise, optical fiber systems, 1843 laser sources, 1776–81 absorption in, 1776 active and confinement layers in, 1777 amplification in, 1776-77

antireflection coatings and, 1779 bandgap in, 1777, 1778 blue violet, in optical memories, 1739 carbon dioxide, 1853 CDROM and, 1735 chirp in, 1844 confinement factor in, 1777, 1778 continuous wave, 1852–53 coupling optics in, 1781 difference frequency generation, 1853 diffraction gratings in, 1779 distributed Bragg reflector, 1780-81, 1780 distributed feedback, 1762, 1779, 1779, 1780, 1853-54 drive electronics in, 1781 effective index in, 1777 electro absorption modulated lasers in, 1826 emission in, 1776 energy bands in, 1777 energy states in, 1776 eye safety in, 1864-65 facet loss in, 1778-79 feedback in, 1776-77 fiber ring, 1771 free space optics in (see free space optics) GaAs lasers, 1853 gain in, 1778 gain switched, 1771 gas type, 1777 heterostructure in, 1777, 1777 history and development of, 1775 holographic memory/optical storage and, 2134 holographic systems and, 1740, 1740 InGaAs, 1853 laser intensity noise in, 1843 laser phase noise in, 1843 lidar, 1863 metal organic vapor phase epitaxy process in, 1778 millimeter wave propagation and vs., 1449 mirrors in, 1776-77 mode locked, 1762, 1961 modenumber and modespacing in, 1777 modulation in, 1779 modulators for, 1781 monitors for, 1781 Nd/YAG, 1853 optical fiber systems and, 1708, 1714, 1842 optical isolators in, 1781 optical memories and, 1739 optical modulators and, 1741-48 optical parametric oscillator in, 1853 packaging and modules in, 1781, 1781 Planck's constant and, 1776, 1777 population inversion in, 1776, 1776 pump type, 1781 pumping in, 1778, 1781 Q-switched, 1762, 1762 quantum cascade, 1853 quasi-Fermi levels in, 1777 rate equations for, 1778 refractive index in, 1779 resonant frequency in, 1779 semiconductor optical amplifiers and, 1781 semiconductor type, 1777-78, 1777 signal quality monitoring and, 2273 single frequency type, 1779 solitons and, 1764, 1770 spontaneous emission in, 1776 spotsize in, 1777 stimulated emission in, 1776 thermoelectric elements in, 1781 threshold current in, 1778 tunable, 1780-81 Vernier effect and, 1780 vertical cavity surface emitting lasers in, 1739, 1781, 1853 wavelength division multiplexing and, 1779 wavelength-selectable, 1779-80, 1780 last mile communications, community antenna TV, 512 last mile technology, powerline communications and as, 1997–98

autocorrelation method in LS, 1262

latency, 549, 1321 latitude, in radiowave propagation, 2064 lattice construction, in low density parity check coding, 662-663, 662, 663 lattice filter, 81 lattice vector quantization, in compression, 644 lattice vector quantization, 2127-28 launch of satellite, spacecraft used, 1251–52 layer 2 forwarding, 2808 layer 2 tunneling protocol, 273, 1651, 2809 layer 2.5 architecture, MPLS, 1594 layer 3 protocols, IP networks, 269 layer 3 signaling, cdma2000, 365-366 layered architecture, in ATM, 200-201 LBG algorithm, compression, 643-644 leaky bucket algorithm, 201, 205, 205, 1659 leaky SAW waves, 2444 leaky wave millimeter wave antenna, 1235-47, 1235, 1428, 1428, 1428 aperture in, 1238 applications and properties of, 1235-36 beamwidth in, 1239 characterization of, 1237 design procedures for, 1239-41 dielectric type, 1244-45, 1244, 1245 dispersion in, 1237, **1237** feeds for, 1245 frequency range for, 1237 interpreting behavior of, 1237-39 layered dielectric guide, 1244-45, 1245 losses in, 1236, 1245 manufacture of, 1245 measurement techniques for, 1245 microstrip, 1241-43, 1243 microwave transmission and, 1236 millimeter wave transmission and, 1236 modulation in, 1241 nonradiative dielectric, 1243-44, 1244 open structures and, 1236 pencil beam radiation using, 1240 phase and leakage constants in, 1237 radiation patterns in, 1235, 1239, 1240-41 scanning properties of, 1235-36, 1239-40 stepped design in, 1241 surface vs. space wave in, 1237 tapering in, 1241, 1242 waveguide and, 1235-36 waveguide and, partially open metallic, 1241 learning in neural networks, 1675, 1677-79, 1677 learning algorithms, in adaptive receivers for spreadspectrum system, 103 least loaded routing, in routing and wavelength assignment in WDM, 2102 least mean square algorithm adaptive antenna arrays and, 69, 71-73 adaptive receivers for spread-spectrum system and, 100-101 blind multiuser detection and, 300-301 cable modems and, 330 in channel modeling, estimation, tracking, 404, 412, 414-415 equalizers and, 83-84, 85, 88, 90 equalizers, 286 packet rate adaptive mobile receivers and 1886, 1883, 1887 polarization mode dispersion and, 1974 in underwater acoustic communications, 41, 44 least squares algorithm in channel modeling, estimation, tracking, 398 equalizers and, 81-82, 84-85 linear predictive coding and, minimization in, 1261-62 least squares smoothing algorithm, in channel modeling, estimation, tracking, 407 Lebesgue decomposition, in pulse position modulation, 2035, **2037**, **2038** Leech lattice, Golay coding, 886 Legendre linear antenna arrays, 148 Lempel-Ziv coding compression and, 638-639

image and video coding and, 1032-33

rate distortion theory and, 2076

lens antenna, 180, 1425-26, 1426, 1427, 2082 LEO satellite networks (see low earth orbit satellite communications) Levinson-Durbin algorithm, 1263, 2349 Levy motion models, in traffic modeling, 1670 lidar lasers, 1863 light emitting diodes (see also optical sources), 1775-76 optical fiber and, sources for, 1708 optical fiber and, 1714 free space optics and (see free space optics) history and development of, 1775 light sources, for optical fiber, 1714 light splitting, in routing and wavelength assignment in WDM, sparse, 2105 light trees, in routing and wavelength assignment in WDM, 2100, 2104 Lightning network, 1720 lightpath topologies in optical cross connects/switches, 1798 routing and wavelength assignment in WDM and, 2098, 2101 lightwave systems for optical fiber, 1707 lightweight directory access protocol, 1656, 1656 likelihood function (see also maximum likelihood estimation), 289, 1338-39, 2026 Lincoln Laboratory Experimental Satellite, 483, 484, 1233 line coding, in partial response signals, 1933-34, 1933 line of sight communications, 208 atrmospheric refraction and, 210-211, 210 broadband wireless access and, 318 diffraction in, 213–215, **214**, **215**, 215–216 indoor propagation models for, 2012-21 local multipoint distribution service and, 318 microwave and, 2555-72 millimeter wave propagation and, 1443-45 path loss and, 1939, 1941 wireless infrared communications and, 2925 line sources and distributions, antenna arrays, 154 line spectral frequencies, speech coding/synthesis, 2350, 2372 line spectral pairs, speech coding/synthesis, 2350, 2821 linear equalizers, 286 linear adaptive equalizers (see also adaptive equalizers), 82-87, 82 linear antenna, 1257-60 beverage antenna and, 1259 dipoles, 1257-58, 1257 directivity in, 1258 gain in, 1258 impedance, impedance matching in, 1258 radiation patterns in, 1257-58, 1258, 1259, 1259 receiving antenna as, 1260, 1260 traveling wave, 1258-60, 1259 linear array active antenna, 62-63, 62, 144-148, 144 microstrip/microstrip patch antenna and, parallel and serial fed, 1373-85, 1373-76, 1381-83 multibeam phased arrays and, single and multibeam, 1514–15, **1515**, **1516** linear block coding, in product coding, 2007 linear broadside antenna arrays, 142 linear coding, 225, 469 linear congrential algorithm, in random number generation, 2292 linear density hard disk drives, 1321 linear detectors for wireless multiuser communications systems, 1616-17 linear equalizers, intapped delay line equalizers, 1688, 1691-92 linear feedback shift register adaptive equalizers and, 85 Bluetooth and, security, 316 cyclic coding and, 619-620, 625-626, 625, 626 linear interpolation, in channel modeling, estimation, tracking, 414, 414 linear minimum probability of error receivers, adaptive receivers for spread-spectrum system, 101-101 linear prediction, 1261-68, 2820-23, 2821, 2822 linear prediction coders, 1261-68 analysis, 1262 applications for, 1264-67

coding excited linear prediction and, 1266-67, 1302-05, 1303 computation of prediction parameters in, 1263-64 encoding/decoding, 1264 example of, 1264 formulation of, 1261-63 least squares minimization in, 1261-62 Levinson-Durbin method in, 1263 linear preditive coding minimum error method in, 1262-63 mixed excitation linear prediction in, 1266-67, 1300, 1306 pulse coding modulation and, 1264 signal to noise ratio and, 1264 speech synthesis and, 1264-67, 1264, 1265, 1266, **1267**, 1300, 2341, 2344–50, **2344**, 2372, 2373 synthesis of, 1263 Toeplitz symmetric matrix in, 1263 in underwater acoustic communications, 37 linear receivers, in adaptive receivers for spread-spectrum system, 98 linear recurring sequences, feedback shift registers, 790 linear time invariant systems, digital filters, 689-90 linear time invariant coders, speech coding/synthesis, 2341 linear time invariant components, simulation, 2287-88 linear time varying components, simulation, 2288-89 linear unequal error protection coding, 2763-69 linearly bounded arrival process, 1568 link access control, cdma2000, 359, 364-365, 365 link access protocol-modem protocol, 1496 link adaptation, multiple input/multiple output systems, 1455 link aggregation, Ethernet, 1284 link budget community antenna TV and, AM systems, 518-519, 522 satellite communications and, 883-884, 1229 ultrawideband radio and, 2760 link layer, TCP/IP model, 541 link layer security, 1153 link layer specific encapsulation, in multiprotocol label switching, 1594 link manager, in Bluetooth, 314 link manager protocol, in Bluetooth, 310, 314 link rerouting, in failure and fault detection/recovery, 1633–34, 1634 Linkabit 268 links, for shallow water acoustic networks, 2206 liquid crystal modulators, in optical synchronous CDMA systems, 1817 liquid crystal switches, 1790-91 liquid crystal optical crossconnects, 1704-05 LiteMAC protocol, 1553 lithium niobate, 1741-48 lithium tantalite, 1742 live backup, 1634 Lloyd's condition and algorithm, in scalar quantization, 2125 Lloyd-Max quantizers, for compression, 642 load, in traffic engineering, 488 load sharing, radio resource management, 2093 loading, orthogonal frequency division multiplexing, adaptive, 1878, 1878 loading elements, in active antenna, 64-65, 64 lobes, antenna, 184 local area networks, 547, 1279-89, 2461 10Base2, 1283 10Base5, 1283 10BaseT, 1283, 1283 100BaseT, 1283-84 1000Base (see Gigabit Ethernet) addressing in, 1282 ALOHA protocols and, 1720 asynchronous transfer mode and, 1719 automatic repeat request and, 224-231 BISDN and, 271 broadcast domains in, 1281 burst mode in, 1284 bus topologies and, 1716, 1716

local area networks (continued) carrier sense multiple access and, 345, 1280-81 coaxial cable for, 1283 code division multiple access and, 458 collision domains in, 1281 collisions in, 1280 copper media for, 1283 dense WDM in, 1720-21 design considerations for, 1717-18 dual queue dual bus and, 1715 dynamic range in, optical, 1718 Ethernet and, 1279, 1501, 1512, 1717, 1719, 1280-81 extranets and, 1163-72, 1165 fault tolerance and, 1639, 1640-42, 1642 fiber distributed data interface and, 1284, 1715, 1718-19 Fibre Channel and, 1719 full duplex Ethernet for, 1284 future of, 1289 Gigabit Ethernet in, 1284, 1721 HiperLAN and, 320-321 history and development of, 1279-80 IEEE 802 standards and, 1280, 1281, 1281-84 intranets and, 1163-72, 1165 ISO reference model and, 1281 jamming in, 1281, 1283 logical link control layer in, 1281-82 MAC addresses in, 1282 media access control and, 1342-49, 1716 media access control frame format in, 1282, 1282 media access control layer in, 1281-82 minimum spanning tree and, 1639-40 multicasting and, 1529-30, 1531-32 multiple link service access points in, 1281 optical bypass in, 1715, 1716 optical fiber and, 1714-22, 1808 optical fiber systems and, 1840 organizationally unique identifiers in, 1282 passive optical networks and, 1717 physical layer in, 1282 point to point communications and, 339, 339 powerline communications and, 1998 protocols and, 1718-19 pulse position modulation and, 2041 reliability and, 1639, 1640-42, 1642 ring topologies and, 1716, 1716 self-healing ring topologies and, 1716, 1716 shallow water acoustic networks and, 2212 slot time in, 1281 software radio and, 2307 star topologies and, 1716-17, 1717 start frame delimiters for, 1282 topologies for, 1715-17 virtual (see virtual LANs) wavelength division multiplexing and, 1719-21, 1720, 2841-42, 2842 wireless infrared communications and, 2925 wireless LAN (see wireless LAN) wireless multiuser communications systems and, 1602 wireless packet data and, 2982 Xerox PARC and, 1279 local decision point, in admission control, 115-116 local exchange carrier and IP telephony, 1177 local gradients, in neural networks, 1678 local loop, wireless, 2947-59, 2948 local multipoint communication services, 1268 local multipoint distribution service, 1268-69 attenuation in, 1273, 1276-77, 1276 bandwidth in, 318, 1268 base station in, 318-319 beamwidth and, 1273 broadband and, 2655, 2671 broadband wireless access and, 317, 318, 322 cell configuration in, 318-319, 319 coverage area of, 1273, 1274, 1273, 1275-76, 1275, 1276 cross polarization discrimination in, 1277, 1277 digital audio-visual council and, 318, 320 digital video broadcasting project and, 318, 320

frequency coordination and interference control in, 1268 frequency for, 1268-69, 1269, 1270 interference and, 318-319 measurement procedures for, 1274-75, 1274, 1275 media access control and, 1269 metropolitan area networks and, 1268 millimeter wave propagation and, 1270-72 multipath interference in, 1273, 1277 quadrature amplitude modulation and, 319, 320 quadrature phase shift keying and, 318, 319, 320 quality of service and, 1269-70 radio channel for, 1272-77 receivers and transmitters for, 1268 regulatory and standards overview for, 1268-70 signal to interference ratio in, 319 standards for, 319-320 terrain attenuation/blockage in, 1273-74 time division multiple access and, 318, 320 time division multiplexing and, 318, 320 transmission loss in, 1275 wireless LAN and, 1269 local optimum in quantization, 2129 local oscillators, 1478 local scaling components, in traffic modeling, 1670 locally optimum Bayes detector, impulsive noise, 2412, 2413 location aided routing, ad hoc wireless networks, 2890 location registration, satellite communications, 1253-54 location, wireless, 2994-95 locator fields, in BCH (nonbinary) and Reed-Solomon coding, 253 locked beam active antenna, 63-66 log a posteriori probability iterative decoding, 561 log periodic antenna, 169, 187 log reflection coefficients, in serially concatenated coding, 2175 logarithmic likelihood function, in maximum likelihood estimation, 1, 1339 logarithmic likelihood ratio, 561 cochannel interference and, 456 multidimensional coding and, 1541-42 serially concatenated coding and, 2168-72 turbo coding and, 2713 logarithmic maximum a posteriori algorithm, soft output decoding algorithms, 2301-02 logical channels, cdma2000, 363-364 logical link control , 546, 547, 1281-82 powerline communications and, 2002 shallow water acoustic networks and, 2208 wireless packet data and, 2983 logical link control and adaptation protocol, 310, 314 logical or virtual topologies, in routing and wavelength assignment in WDM, 2100-01 long range dependent models, in traffic modeling, 1667-70 long term prediction, in speech coding/synthesis, 2823-25 long waves (see low frequency) long wire antenna, 180, 188 longitudinal saturation recording, 1323 Longley Rice model, radiowave propagation, 216 look ahead encoder, magnetic recording systems, 2254 look ahead maximize batch, 235 lookahead, constrained coding techniques for data storage, 578 lookup table, in predistortion/compensation in RF power amplifiers, predistortion, 533-534, 533 loop antenna arrays, 142 loop antenna, 1290-99 analysis of, 1290–96 applications for, 1290, 1296-98 body worn loops, 1297, 1297 boundary condition matching in, 1294-95 current density, surface, 1294 directivity in, 1293-94 far field (Fraunhofer zone) in, 1292 fat wire, 1294, 1294 ferrite loaded, 1296-97, 1296 gap fed, 1294, 1295-96 high frequency small resonanted, 1297-98

impedance, wave impedance in, 1292, 1292, 1292-95 induction zone in, 1292-93, 1293 infinitesimal, 1290-94, 1290 intermediate field in, 1293 Lorentz condition in, 1291 magnetic current/field in, 1291-92 near field (Fresnel zone), 1292, 1293 quad, 1298 radiation patterns in, 1292 radiation resistance in, 1295 reactance in, 1295 rectangular, 1298, 1298 television and FM broadcasting, 2517-36 vector and scalar potentials in, 1290-91, 1294 wire used in, NEC rating for, 1296 loop delay, in power control, 1986 loop filter, frequency synthesizers, 845, 846-848 loop gain, frequency synthesizers, 851–853, **852**, **853** loopback calls, ATM, 207 loopback, 1636, 1636 Lorentz condition, in loop antenna, 1291 Lorentzian transition response, in digital magnetic recording channel, 1324-25, 1324, 1328-29, 1329 Lorenz sequence, in chaotic systems, 429-430, 429 loss control circuit, acoustic echo cancellation, 1, 2 loss resistance, antenna, 184 lossless compression, 6320639, 2123, 2124 lossless source coding, 2069 lossy coding, in speech coding/synthesis, 2341 lossy compression, 371, 632-633, 639-648, 2123, 2124 LOTRAN database and calculations, 1856-57 loudspeaker transducers (acoustic), 34 loudspeaker enclosure microphone system, 1-3, 2, 6 low bit rate speech coding, 1299-08 algebraic CELP in, 1304, 1306 algebraic vector quantized CELP in, 1306 analysis by synthesis method in, 1302-03 channel coding and, 1299 characteristic waveforms in, 1301-02, 1302 coding excited linear prediction in, 1302-05, 1303 conjugate structure CELP in, 1304, 1306 enhanced full rate coders in, 1306 enhanced variable rate coder in, 1306 focused search technique in, 1306 joint position and amplitude search in, 1306 mean opinion score in, 1305 mixed excitation linear prediction and, 1300, 1306 modeling for, 1300-02 parameter estimation from speech segments in, 1302-05 perceptual speech quality measure in, 1305 pitch synchronous innovation CELP in, 1304 predictive coding for, 1300 pulse coding modulation and, 1299 rapidly evolving waveform in, 1301 regular pulse excitation with long term predictor in, 1304 signal to noise ratio in, 1305 sinusoidal coders for, 1300-01 slowly evolving waveform in, 1301-02, 1302 text to speech systems in, 1304-05 waveform interpolation coding for, 1301-02 low delay CELP, in speech coding/synthesis, 2349, 2355, 2825–26, **2826** low density parity check coding, 658-668, 802, 1308-18 a posteriori probability decoders in, 1313, 1316 additive white Gaussian noise and, 658, 1312, 1313 algebraic type, 1316 applications for, 663-65 balanced incomplete block design in, 658, 659-661, 659, 1316 BCJR algorithm and, 1316 binary symmetric channel for, 1315 bit error rate in, 1309, 1309, 1316 bit flipping in, 1311-12, 1312 Buratti construction in, 661 check bits and, 1308 combinatorial design and, 1316 cycles in, 1311 density evolution in, 1315 designing, 1315-16

low density parity check coding (continued) encoding in, 1316-17 error correcting coding in, 1316 Euclidean distance in, 661-662 factor graphs in, 1316 future of, 1316-17 Gallagher coding in, 658, 659 Galois fields in, 661 generator matrix in, 1310 Golay coding and, 659 Hamming distance in, 1309, 1310 hard vs. soft decision in, 1309, 1312 hyperplanes in, 661 iterative coding and, 1309 iterative decoding in, 1311-12, 1316 Kirkman system in, 659, 664-65 lattice construction in, 662-663, 662, 663 low density coding in, 1310-11 magnetic recording systems and, 2266-67 marginalize product of functions problem and, 1316 maximum likelihood decoding and, 1317 min-sum algorithm in, 1315 Netto's constructions in, 661 parity check coding in, 1309-10 parity check matrix in, 1310 product coding and, 2011 regular vs. irregular, 1310 Shannon or channel capacity and, 1308 simulations of, 664-65 state variables and, 1316 Steiner triple system in, 659 sum product decoding in, 1309, 1312-15, 1314 Tanner graph of, 1311, 1311, 1312 turbo coding and, 658, 1312, 1316 vertices in, bit or parity check, 1311 wireless and, 1316 wireless multiuser communications systems and, 1610 low earth orbit, 196, 196, 1223-24, 1231-32, 1247-56, 2112, 2119 additive white Gaussian noise and, 1251 ALOHA protocols in, 1253 apogee and perigee in orbit of, 1248 applications for (Iridium, Globalstar), 1247 ascending node in orbit of, 1248 Big Leo systems in, 1251 broadcast satellite service in, 1251 C band, 1251 circuit switched network architectures and, 1253-54 constellation of, 1247, 1248, 1249-50 coverage or footprint in, 1249 feeder links for, 1251 fixed satellite service in, 1251 frequencies used in, 1251 future of, 1255-56, 1255 geostationary satellite in, 1248, 1250-52 global positioning system and, 1254 Globalstar in, 1251 handoffs in, 1252, 1254 interference and, 1251 Internet protocol and, 1253 intersatellite links in, 1252 Iridium and, 1251, 1253 Ka band, 1251 Ku band, 1251 L band, 1251 launch of, spacecraft used for, 1251-52 link performance in, 1251 location registration in, 1253-54 mobile satellite service in, 1251 Molnya orbit in, 1250 multihop satellite routing in, 1254 multiple access in, 1253 networking considerations in, 1252-55 Orbcomm in, 1251 orbital geometry for, 1248-49, 1248 packet switched architectures and, 1255 propagation delay and, 1250–51 retrograde orbits and, 1248 seams in orbits of, 1250 spot beams in, 1249

station keeping in, 1248 subscriber links for, 1251 time division multiple access in, 1253 tracking in, 1252 Tundra orbit in, 1250 very small aperture terminal and, 1247 Walker delta or rosette constellatioin in, 1250, 1250 Walker star or polar constellations in, 1250, **125** low frequency, 2059–69 low noise amplifier, 327 low probability of intercept, in chaotic systems, 428 lower sideband amplitude modulation, 133 lower-upper decomposition, in antenna modeling, 173 lowpass pulse amplitude modulation, 2022 lowpass filter, 134, 136, 414 lowpass signals discrete time representation of, 2286 random processes, sampling of, 2286-87 sampling and, 2109-10, 2109 Luke polyphase sequences, 1979 Lyapunov exponents, in chaotic systems, 429-430 M algorithm, in adaptive equalizers, 81 M-ary phase shift keying, 710-711, 1335, 1976 M-ary time shift keying, 1335 MAC addresses Bluetooth and, 312, 315 Ethernet and, 1503 local area networks and, 1282 multicasting and, 1529, 1531-32, 1532 security and, 1646 MAC frames, Ethernet, 1502-03, 1503 MAC sublayer Ethernet and, 1502 powerline communications and, 2002, 2003-04 Mach-Zehnder filters/interferometers, 1723-24, 1723, 1730, 1730 optical fiber and, 1709 optical multiplexing and demultiplexing and, 1749 optical couplers and, 1698–99 optical cross connects/switches and, 1785 optical filters and, 1757-58, 1757 optical modulators and, 1742-44, 1743 optical signal regeneration and, 1760-61, 1760 signal quality monitoring and, 2273 optical transceivers and, 1826-27, 1826 photonic analog to digital conversion and, 1961-64 macrobending attenuation, optical fiber, 439 macrocells, 376, 449, **450**, 1940–41 macroflows, in flow control, traffic management, 1568, 1653-55 magentoresistive materials, 2249 magnetic coordinates, radiowave propagation, 2061 magnetic field, antenna, 171, 180 loop antenna and, 1291-92 radiowave propagation and, 2063-64 waveguide and, 1394 magnetic flux, active antenna, 55 magnetic recording systems additive white Gaussian noise, 2253, 2259, 2261, 2262, 2264 bit error rate in, 473, 2266, 2266 block coding in, 2257 block error rate in, 473 block missynchronization detection in, 471-472 Butterworth filters and, 2262 coding channels for, 466-476 combined modulation/parity coding, 2257-58 communications channels in, 2249-51 cyclic coding in, 469 data dependent NPML detection in, 2265-66 decision feedback equalizer in, 2262-63 equalization and detection in, 2258-63 error correcting coding in, 466-467, 466, 470, 472-474 error detecting coding in, separate vs. embedded, 474 error detection and correction in, 2256-58, 2257 error rate definitions for, 473 Euclidean distance and, 2249, 2260 filtering in, 2262 finite impulse response filters and, 2259, 2262 finite state transition diagram in, 2253-57, 2256

frequency response in, 2251-52, 2252 future trends in, 2266-67 Gaussian-Markov noise in, 2265-66, 2265 hard and soft decision in, 475, 2257 history and development of, 2247-48 interleaving vs. noninterleaving in, 472 intersymbol interference and, 2251-52, 2251 linear coding in, 469 look ahead encoder in, 2254 Lorentzian pulse in, 2250, 2250 low density parity check coding in, 2266-67 magentoresistive materials in, 2249 matched filters and, 2262 maximum likelihood sequence detection in, 2258-60, 2263 maximum transition run in, 2248, 2249, 2253, 2255-58, 2256 microtrack model for, 2252-53, 2252 minimum mean square error, 2261 modulation in, 2249-51, 2253-58 noise and, 2257 noise predictive maximum likelihood and, 2248, 2250, 2261–66, **2261**, **2262** non return to zero in, 2250, 2250 non return to zero inverted in, 2250, 2250 Nyquist frequency in, 2261 parity-based post processing in, 2263-65, 2264 partial response shaping in, 2248 partial response maximum likelihood in, 2248, 2253-55, 2258, 2259-65, 2260 performance and 472-474 post processing in, 2263-65, 2264 redundant array of independent disks and, 474-475 Reed-Solomon coding and, 467-475, 2249 run length limited in, 2248, 2249, 2254 saturation recording in, 2248-49, 2249 serially concatenated coding and, 2175-76, 2175 signal processing in, 2247-68 signal to noise ratio, 2264 sliding block decoder in, 2254 soft bit error rate in, 474 soft decision decoding algorithms for, 475 state transition diagrams in, 2253-57, 2256 symbol error rate in, 473 systematic coding in, 469 tape drive ECC and, 474 tracks in, 2248 trellis coding and, 2260-61, 2260 turbo coding and, 2266-67 variable gain amplifier in, 2250 Viterbi algorithm and, 2259, 2260, 2265 write process in, 2249 magnetic storage systems (see also hard disk drives), 1319-34 additive white Gaussian noise and, 1332 analog to digital conversion in, 1319 archival systems and, 1319 backup systems and, 1319 bandwidth and, 1326 channel coding in, 1331-33 channel identification in, 1326 compact disk read only memory and, 1319 cost per megabyte of, 1319 digital audio tape in, 1319 digital forms of, 1319 digital magnetic recording channel in, 1322-26, 1326, 1327 distortion in, 1325-26 dropouts in, 1326 error correction coding and, 1326 error detection and correction in, 1332 extended partial response in, 1328-32, 1329, 1331 filters in, 1329-30, 1333 finite impulse response equalizers and, 1324, 1329 frequency modulation and, 1327 generalized partial response in, 1332 generalized partial response in, 1331-33 hard disk drives in, 1319, 1320-22 head noise in, 1325 interleaving in, 2, 1330 intersymbol interference and, 1325-31

magnetic storage systems (see also hard disk drives) (continued) intertrack interference (crosstalk) in, 1325 JPEG compression and, 1319 longitudinal saturation recording and, 1323 Lorentzian transition response in, 1324-25, 1324, 1328-29, 1329 maximum likelihood sequence detector in, 1331 mechanical memories in, 1319 media noise in, 1325 M-H curve in, 1322-26, 1326 minimum mean squared error and, 1, 1329 modified frequency modulation in, 1327 modulation coding and, 1326 MP3 audio compression and, 1319 MPEG compression and, 1319 non return to zero inverse and, 1327 non return to zero inverse interleaved, 1330 normalized linear density in, 1324 Nyquist frequencies in, 2, 1330 parity coding in, 1331-33 partial erasure in, 1326 partial response maximum likelihood in, 1328, 1328, 1330–31 peak detection in, 1327 peak detection in, 4, 1332 performance of, 1319 personal video recorder and, 1319 preamplifier noise in, 1325 pulse amplituide modulation in, 1323 pulse coding modulation and, 1319 RAMAC systems in, 1320, 1321 read process in, 1320, 1323-28 redundant array of independent disks and, 1322 Reed-Solomon coding and, 1326 run length limited coding in, 1327 segmentation of, 1319 signal processing and coding in, 1326-33, 1327 signal to noise ratio and, 1326, 1327, 1331 solid state memories in, 1319 storage devices and, 1319 stripe in, magnetoresistive, 1323 thermal asperity in, 1326 time varying maximum transition run length coding in. 1332 transition shift in, 1325, 1325 trellis coding in, 1331-33 turbo coding in, 1326-27 Viterbi detectors and, 1330-33 write once read many devices and, 1319 write process in, 1320, 1323-26 magnetooptic disks, 1319, 1738-40 magnetooptic magnetic field modulation, 1739 magnetron antenna, 179 magnification, parabolic and reflector antenna, 1922 magnitude, antenna arrays, 144, 145 main (major) lobe antenna, 184 man in the middle attacks, 611, 2810 man machine interface, wireless application protocol, 2901 management informatioin base, ATM, 200 management plane, ATM, 264 mandatory access control, 1649 Mandelbrot, 421 manipulation detection coding, 613 map symbol-by-map symbol, 89, 89 mapping, image and video coding, 1030 Marcelling-Fischer coding, in compression, 646 Marconi, Guglielmo, 179, 188, 677, 1477, 2585 marginalize product of functions problem, low density parity check coding, 1316 Marisat satellite communications, 876 maritime communication systems, 1434, 1477 mark edge noise, constrained coding techniques for data storage, 573 Markov chain, 963 statistical multiplexing and, 2423-32 traffic engineering and, 492 Markov modulated Bernoulli process, 117 Markov modulated fluid process, 117 Markov modulated models, in traffic modeling, 1668

Markov modulated Poisson process, 117, 1668, 1671, 2423, 2424 Markov source, in compression, 632, 634, 1033 Markov/semi-Markov models, 1666-68, 2291, 2291 Marsaglia-Zamann algorithm, random number generation, 2292 MASCARA protocol, 2908 masked threshold, in speech coding/synthesis, 2364 masking digital audio broadcasting and, 682, 682 speech coding/synthesis and, 2364 masking spectrum, in speech coding/synthesis, 2345 Massachussetts Institute of Technology, 267 Massey-Berlekamp decoding algorithm, 257-260, 470, 617, 625-626 master slave configuration, Bluetooth, 314-316 mastering of discs, CDROM, 1734 matched coding, serially concatenated coding for CPM, 2180 matched filters, 700, 1116, 1335-38, 1336 additive white Gaussian noise and, 1336-37 chirp modulation and, 442-443, 443, 446 frequency domain interpretation in, 1337 magnetic recording systems and, 2262 properties of, 1336-37 pulse amplitude modulation and, 2026, 2029, 2030 signal to noise ratio and, 1337 signature sequence for CDMA and, 2275 tapped delay line equalizers and, 1691 tropospheric scatter communications and, 2700, 2700 wireless multiuser communications systems and, 1616 matched nodes, SONET, 1637 matching networks, waveguide, 1409-11, 1409, 1410 material dispersion, in optical fiber, 1711 material properties, in indoor propagation models, 2013-14 Max-Batch batching, 235 maximal length signature sequence for CDMA, 2279-81. 2280 maximal ratio combining, 731 fading and, 788 multicarrier CDMA and, 1527 quadrature amplitude modulation and, 2051-52, 2051 satellite communications and, 1230 spatiotemporal signal processing and, 2336 wireless and, 2920 wireless multiuser communications systems and, 1619 maximum a posteriori algorithm adaptive equalizers and, 79, 81, 89 convolutional coding and, 600 demodulation and, 7, 1335 iterative decoding and, 561 maximum likelihood estimation and, 1341 minimum shift keying and, 1472, 1473-74 multidimensional coding and, 1542 pulse amplitude modulation and, 2026 soft output decoding algorithms and, 2295, 2297, 2299–2301, 2299 space-time coding and, 2328 turbo coding and, 2705, 2714 turbo trellis coded modulation and, 2743-45, 2750 wireless multiuser communications systems and, 1616 maximum APP sequence detection, serially concatenated coding for CPM, 2182 maximum burst size, 117, 266, 551, 1656, 1658 maximum cell transfer delay, ATM, 266 maximum coding, in BCH (nonbinary) and Reed-Solomon coding, 254 maximum distance separable coding, in BCH (nonbinary) and Reed-Solomon coding, 254 maximum factor queue length batching, 234-235 maximum laxity first scheduling, medium access control, 1555 maximum length signal generator, adaptive equalizers, 85 maximum likelihood algorithm

bit interleaved coded modulation and, 277, 283 blind equalizers and, 289–291

in channel modeling, estimation, tracking, 402-405, 427 concatenated convolutional coding and, 558-560 convolutional coding and, 600 cyclic coding and, 620 equalizers and, 89-91 expectation maximization algorithm and, 769-780 hidden Markov models and, 961-962 low density parity check coding and, 1317 maximum likelihood estimation and, 1341-42 multicarrier CDMA and, 1527 multiple input/multiple output systems and, 1455 permutation coding and, 1953, 1954 pulse amplitude modulation and, 2026 sequential decoding of convolutional coding and, 2143, 2145–47 serially concatenated coding and, 2164, 2167-68 space-time coding and, 2325, 2327 trellis coding and, 2638, 2646-48 tropospheric scatter communications and, 2702-03 maximum likelihood estimation, 1338-42 asymptotic properties of estimators using, 1339-40 bias and variance in, 1339 bounds in, 1339 Byes estimation of random parameter in, 2, 1340 in channel modeling, estimation, tracking, 398 conditional mean estimator using, 1340-41, 1340 cost functions in, 1340 Cramer-Rao inequality and bound in, 1339 expectation maximization algorithm and, 1341 Fisher's information matrix in, 1339, 1340 hidden Markov models and, 1341 likelihood function in, 1338-39 loglikelihood function in, 1339 maximum a posteriori algorithm, 1341 maximum likelihood and, 1341-42 minimum mean square error and, 1341 monotonic transformation and, 1339 properties of estimators using, 1339-40 score and scores vector in, 1339 maximum likelihood sequence algorithm, 286, 1331, 2258-60, 2263 maximum likelihood sequence estimation adaptive equalizers and, 79, 81, 90 blind equalizers and, 297 in channel modeling, estimation, tracking, 417-418 cochannel interference and, 455 minimum shift keying and, 1458, 1469-70, 1469 multiple input/multiple output systems and, 1455 partial response signals and, 1932, 1933 polarization mode dispersion and, 1974 space-time coding and, 2328, 2329 Viterbi algorithm and, 2817–18 maximum likelihood sequential decoding algorithm, 2140, 2155-56, 2182-87 maximum logarithmic MAP algorithm, soft output decoding algorithms, 2302 maximum mutual information, hidden Markov models, 962 maximum queue length batching, 234-235 maximum reuse routing, routing and wavelength assignment in WDM, 2102-03 maximum segment size (MSS), transmission control protocol, 2604, 2606 maximum transition run, 581-582, 2248, 2249, 2253-58, 2256 maximum usable frequency, 949, 2065, 2066, 2067 max-min fairness, flow control, traffic management, 1653 Maxwell's equations active antenna and, 53-54 adaptive antenna arrays and, 70 antenna modeling and, 169-172, 176, 179-181 path loss and, 1936 waveguide and, 1390 M-band filters orthogonal transmultiplexers, 1882-83, 1882, 1884 MD5, 218 mean cost function, blind equalizers, 291 mean distance ordered partial search, vector quantization, 2126

memory

mean effective gain, antenna for mobile communications, 192-193 mean opinion score, in speech coding/synthesis, 1179, 1305, 2352, **2353**, 2819–20 mean removed vector quantization, 2127 mean squared error algorithm adaptive antenna arrays and, 187 blind multiuser detection and, 303 in channel modeling, estimation, tracking, 413, 415 equalizers and, 81, 83-84, 84, 86, 88 fading and, 781 rate distortion theory and, 2069-80 spatiotemporal signal processing and, 2338-39, 2339 speech coding/synthesis and, 2347-48 synchronization and, 2475 transform coding and, 2593 in underwater acoustic communications, 42 meander patch antenna, 193, 194, 194 measurement-based admission control, 118, 1656 mechanical memories, 1319 media access control, 15, 547, 549, 1342-49 ADAPT protocol and, 1348 adaptive round robin and earliest available time scheduling in, 1556, 1559 allocation-based protocols in, 1343, 1344-46, 1344 ALOHA protocols and, 1346, 1347, 1552, 1553, 1559 application level data units and, 1553 ATM and, 2907-09 backoff schemes and, 1346 bucket credit weighted algorithms in, 1556 busy tone multiple access and, 1347 cable modems and, 324, 334-335 carrier sense multiple access and, 346, 1346-47 CATA protocol and, 0, 1348 cdma2000 and, 359, 363-365 cellular digital packet data and, 9, 1347 centralized vs. distributed protocols in, 5, 1343 channel available time table in, 1554 classification of protocols using, 1342-44 clear to send in, 1348 code division multiple access and, 1343-45, 1348 collision free protocols for, 1557 connectivity and, 1343 contention-based protocols in, 1343, 1346-47 credit weighted algorithms in, 1556 cyclic reservation multiple access and, 1558 deterministically guaranteed service and, 1556 direct sequence CDMA and, 7, 1345 directional antenna and, 1348 distributed coordinated function in, 1348 distributed queue dual bus and, 1558 distributed queue multiple access in, 1558 dynamic allocation schemes and, 1553 dynamic time wavelength division multiple access and, 1552 earliest available time scheduling in, 1554 energy efficient, 0, 1348 Ethernet and, 1347, 1501 fair distributed queue in, 1558 FairNet and, 1558 FatMAC protocol and, 1553 fiber distributed data interface and, 1346 fixed assignment in, 1552 fixed priority oriented demand assignment in, 1347 flying target algorithm in, 1557 frequency division multiple access and, 1344 frequency domain protocols in, 1342-43 frequency hopping CDMA and, 1344-45, 1345 global system for mobile and, 1343, 1344 graceful degradation and, 1345 graph coloring problem and, 1344 hidden and exposed terminal problems in, 1343, 1347 HIPERLAN and, 1348 hop in, 1344 hybrid domain protocols in, 1343 hybrid optical networks and, 1559 hybrid protocols in, 1347 hybrid TDM and, 1553 IATSAMTR scheduling in, 1558 interframe spacing and, 1347

KTMTR scheduling and, 1558 LiteMAC protocol and, 1553 local area networks and, 1281-82 local multipoint distribution services and, 1269 maximum laxity first scheduling in, 1555 Meta-MAC protocol and, 1348 minimum laxity first with time tolerance scheduling in, 1555-56 MPEG compression and, 1558, 1559 multimedia applications and, 1558-60, 1559 multimedia wavelength division multiple access and, 1558 multiple access collision avoidance and, 1347-48 near far problem in, 1343 nonpretransmission coordination protocols in, 1552 optical networks and, 1551-62, 1716 packet demand assignment multiple access in, 1347 packet transmission and, 1551-55 partial fixed assignment protocols in, 1552 physical link characteristics and, 1343 point coordination function and, 1348 polling in, centralized vs. distributed, 1345-46 powerline communications and, 1995, 2003-04 pretransmission coordination protocols in, 1552-53 priority index algorithm in, 1557 propagation time and, 1343 protocol threading in, 1348 quality of service and, 1556-59 random access protocols in, 1552 random delay and, 1346 real time service and, 1555-58 receiver available time table in, 1554, 1555 receiver oriented earliest available time scheduling in, 1554-55 receiver oriented protocols using, for variable length messages, 1554 reservation protocol and, 1558 reservation-based protocols in, 1343, 1347-48, 1552-54 satellite communications and, 879 sensor networks and, 1348 shallow water acoustic networks and, 2208, 2209-10, 2215-17, 2216 spatial reuse and, 1344 splitting algorithms and, 1347 spread spectrum and, 1343, 1344-45 statistically guaranteed service and, 1556-58 synchronous round robin with reservation in, 1557 throughput and, 1344 time deterministic time/wavelength division multiple access in, 1555 time division multiple access and, 1343, 1344, 1347 time division multiplexing and, 1552, 1553 time division WDMA and, 1553 time domain protocols in, slotted and unslotted, 1342 time spread multiple access in, 1348 token ring and, 1345-46 transmitter scheduling algorithm in, 1557 tree algorithms and, 1347 validated queue algorithm in, 1556 variable length message transmission and, 1551-55 video on demand and, 1558 wavelength division multiple access and, 1558 wavelength division multiplexing and, 1551-52, 1551. 2842 wireless communications, wireless LAN and, 1285-87, 1343, 2942 wireless packet data and, 2982 wireless sensor networks and, 2993-94 media gateways, session initiation protocol, 2198 media independent interface, Ethernet, 1502, 1506, 1507 media noise, digital magnetic recording channel, 1325 medium access unit, Ethernet, 1506 medium dependent interface, Ethernet, 1508, 1509, 1508 medium earth orbit satellite, 196, 196, 1223, 1224, 1231, 1232, 1249, 2112 medium frequency, 208, 2059-60 medium wave, 1477 MEGACO/H.248 session initiation protocol, 2198 mel scale, automatic speech recognition, 2373 mel scale frequency cepstral coefficients, in automatic speech recognition, 2373, 2382

constrained coding techniques for data storage and, 575 minimum shift keying and, 1457, 1458 neural networks and, 1677 optical, 1733-41, 1734 transport protocols for optical networks and, 2620 memory nonlinearities, simulation, 2290 memory order, in sequential decoding of convolutional coding, 2141 memory requirements of quantization, 2128 memoryless modulation pulse amplitude modulation and, 2024 sequential decoding of convolutional coding and, 2144 serially concatenated coding and, 2173 memoryless nonlinearities, in simulation, 2289-90 memoryless source, in compression, 632, 633-634, 641 Menger's theorem, 1635 merging bits, merging rules, constrained coding techniques for data storage, 576 merging edge, trellis coded modulation, 2627 merit factors Golay complementary sequences and, 895 peak to average power ratio and, 1951 mesh networks cycle covers and, 1638-39, 1638 fault tolerance and, 1637-39, 1638, 1639 reliability and, 1637-39, 1638, 1639 message authentication coding, 218, 613 message passing, product coding, 2011 message signal, amplitude modulation, 132, 133 messages, protocols, 538-556 metal organic vapor phase epitaxy process, 1778 metal oxide semiconductor FET active antenna, 57-58, 57, 58 metal semiconductor FET active antenna, 57-58, 57, 58 metal semiconductor metal photodetectors, 1001-02 Meta-MAC protocol, 1348 method of moments adaptive antenna arrays and, 68 antenna arrays and, 165 antenna modeling and, 173, 174, 175 waveguide and, 1420 metropolitan area exchanges, IP networks, 268 metropolitan area networks, 2461 Ethernet and, 1512 fault tolerance and, 1632 HiperMAN and, 320 local multipoint distribution services and, 1268 multicasting and, 1529 optical fiber and, 1714, 1808 optical fiber systems and, 1840 reliability and, 1632 wavelength division multiplexing and, 2862-73, 2863 M-H curve, 1322-26, 1326 Michelson interferometer optical multiplexing and demultiplexing and, 1749 optical signal regeneration and, 1760-61, 1760 microbending sensitivity, optical fiber, 439 microcells, 376, 449, 450, 1941 microelectromechanical systems, 1349-56 cellular digital packet data and, 1350 code division multiple access and, 1350 deep reactive ion etching in, 1352 devices, circuits, systems using, 1352-55 digital European cordless telecommunications and, 1350 fabrication techniques for, fundamentals of, 1350-52, 1351 general packet radio service and, 1350 global system for mobile and, 1350 Internet protocol and, 1350 micromaching in, 1351, 1352 optical cross connects/switches and, 1705, 1705, 1784, 1785, 1793-96, 1794, 1795 optical multiplexing and demultiplexing and, 1758 radio frequency components and, 1350 sacrificial and structural layers in, 1351 stiction in, 1351 wireless sensor networks and, 2990-96

microflow, in flow control, traffic management, 1653 microfluidoptical crossconnects, 1704, 1704 micromaching, microelectromechanical systems, 1351-52, 1352 microphone transducers (acoustic), 34, 34 microstrip end launcher waveguide and, 1400, 1400, 1404, 1405, 1400 waveguide and, 1401–05, **1401**, **1402**, 1401 microstrip E-plan probe, 1399, **1399**, 1399 microstrip feed line, 51-52, 52, 62, 62, 1362-63, 1362 microstrip/microstrip patch or line antenna, 180, 184, 187, 193, 197, **197**, 199, 1356–80, **1357** antenna arrays and, 142, 152, 152 aperture coupled microstrip feed line for, 1362-63, 1363 aperture coupled type, 1368-70, 1371 arrays of, 1371–77, 1380–90 bandwidth in, 1360, 1364-70 blindness in, 1371, 1376, 1388 cavity models of, 1357 coaxial feed for, 1361-62, 1362 comb line arrays in, 1374, 1376 combination feeds in, 1383-84, 1383 configuration and shapes used in, 1361, 1362 coplanar microstrip feed line for, 1362-63, 1362 coplanar parasitic elements in, 1367–68, **1368** design of, 1358–61 development of, 1356-57 dielectric substrate in, 1358, 1360, 1361, 1363, 1365 directivity in, 1360-61 electrical characteristics of, 1357-61, 1357 element length in, 1359, 1359 element width in, 1358-59, 1359 fabrication of, tolerances and impact of, 1361, 1361 FEDCOMA and FEGCOMA configurations in, 1368-70 feed methods for, 1361-63 feeds for, 1373-74, 1373, 1374, 1380, 1383-84, 1383 fixed-beam planar array of, 1386-87, 1386, 1387 gain in, 1360-61, 1361, 1380 half power beamwidth for, 1358 impedance matching in, 1363, 1383 input impedance in, 1358, 1359, 1360, 1362 leaky wave antenna and, 1241-43, 1243 linear arrays of, parallel and serial fed, 1373-85, 1373-76, 1381-83 losses associated with, 1359-60, 1360 microstrip feed line for, 1362-63, 1362 millimeter wave antenna and, 1429, 1429, 1430 multiresonator type, 1367-70 mutual coupling in, 1370-71, 1371, 1372, 1373 narrowband configuration of, 1357 nonresonant methods for bandwidth enhancement in, 1366 - 67phased array of, 1384-85, 1384, 1385 planar arrays of, 1374-75, 1377, 1385-89, 1386, 1387 polarization (linear, circular) properties in, 1363-64, 1363, 1364 printed circuit board for, 1380 pros and cons of, 1357 proximity coupled microstrip feed line for, 1362-63, 1363 quality factor in, 1357, 1359-60, 1364 radiation patterns in, 1357-58, 1358, 1359 reflectarray type, 1387, 1387 resonant frequency/resonant dimension of, 1359, 1359, 1361, 1361 scanning arrays of, 1375-77, 1384-85, 1387-89 single patch bandwidth in, 1365-66, 1365, 1366 spiders in, 1388 stacked type, 1367, 1367 suspended type, 1365-66 transmission line models of, 1357 voltage standing wave ratio in, 1360, 1363, 1366, 1367 waveguide arrays of, 1373-74 Wilkinson dividers for, 1373, 1382 microtrack model, magnetic recording systems, 2252-53, 2252 microwave, 208, 1706, 2179, 2555-72

absorption in, 2560 active antenna and, 47-48, 49, 65 adaptive equalizers and, 2569-70 anomalous propagation in, 2559-60 antenna for, 179, 180, 2567 atmospheric effects and, 2558-60, 2559 attenuation, 2560 automatic gain control and, 2567 bit error rate (BER) in, 2565–67, **2566** broadband wireless access and, 317 channel models for, 2564-65 community antenna TV and, 513 digital elevation models for, 2561 digital radio design and, 2567-70, 2568 digital terrain elevation data and, 2561 dispersion in, 2565 dispersive fade margin in, 2565, 2566 diversity and, 2563-66, 2567-69 equalization and, 2565 fading in, 2562-65, 2562, 2571 free space loss in, 2556 frequency allocation in, 2566-67 Fresnel zones in, 2557-58, 2558 gain in, 2570-71 interference and, 2566-67 inverse bending (earth bulge) in, 2559 land use/land clutter in, 2561 leaky wave antenna and, 1236 line of sight transmission and, 2555-72 multipath in, 2562-65, 2562 non return to zero in, 2567 parabolic and reflector antenna and, 1927-28 path profiles for, 2560-62, 2561 protection systems in, 2569 quadrature amplitude modulation and, 2569, 2570 radio link calculations for, 2570-71 rain attenuation in, 2560 receivers for, 2567-70, 2568 reflection in, 2556-57, 2557 refraction and, 2558-60, 2559 signal to noise ratio, 2567, 2571 subrefraction in, 2559 superrefraction in, 2559-60 terrain effects in, 2556–58 transmitters for, 2567–70, **2568** traveling wave tubes and, 2567 waveguide for, 1390-1423 microwave multipoint distribution service, 317 microwave waveguide (see waveguide) midamble adaptive equalizers and, 90 packet rate adaptive mobile receivers and, 1899 Middleton model, in impulsive noise, 2409-10, 2416 Mie scatter, free space optics, 1855-57 Military Amateur Radio Service, 1478 military applications for antenna, 169 military communications, 1477 military use of wireless LANs, 2680-82 Miller coding, in constrained coding techniques for data storage, 576 Miller Rabin cryptography, 614-615 millimeter wave antennas, 1423-33 active integrated antenna, 1429-31, 1430, 1431 amplitude modulation and, 1425 application of millimeter waves and, 1425 bandwidth in, 1425 beam steering type, 1431, 1432 efficiency in, 1425 frequency and, 1423-25, 1424 gain in, 1425 horn type, 1425, 1425, 1427-28, 1427 leaky wave antenna and, 1236 leaky wave, 1428, 1428 local multipoint distribution services in, 1268-79 losses in, 1425 microstrip/microstrip patch antenna and, 1429, 1429, 1430 millimeter waves defined for, 1423-25, 1424 nonradiative dielectric waveguide and, 1428, 1428 optically controlled, 1431, 1431 periodic dielectric type, 1428, 1428

planar AIA in, 1430-31, 1431 printed circuit type, 1428-29, 1429 radiation patterns in, 1425 reflector and lens type, 1425-26, 1426, 1427 slotted waveguide type, 1428, 1429 waveguide based, 1426-28, 1427, 1428 waveguide derived, 1427-28 wavelength in, 1423 millimeter wave propagation, 1270-72, 1423-25, 1424, 1433-50 absorption and emission in, 1270, 1270, 1436-37, 1439 absorption of, in clear air, 1270, 1270 angle error in, 1435 applications for, 1434 atmospheric effects on, 1434-43 atmospheric particulate effects and, 1439–43 attenuation in, 1443–45, **1444**, 1445–48, **1446**, **1447** bandwidth and, 1434 bending in, 1435 Brewster angle in, 1438 cloud, fog, haze attenuation in, 1442-43, 1443 depolarization in, 1272, 1439-40, 1445 diffraction in, 1438-39, 1445 ducting in, 1435, 1445 Earth-space transmission paths for, 1445 emission in, 1436–37, 1445–46 fog attenuation in, 1272 Fresnel zone and, 1438 global positioning system and, 1436 Kirchhoff's law and, 1437 laser vs., 1449 line of sight transmission in, 1443-45 maritime communication systems using, 1434 multipath interference and, 1434, 1445 multipath interference in, 1438 oxygen and absorption in, 1437, 1437 propagation effects on, 1433-34 radio relay systems using, 1434 rain attenuation and, 1270-72, 1440-45, 1440, 1441 Rayleigh scattering in, 1271 refraction and refractive index in, 1434-36, 1435, 1445 sand and dust attenuation, 1443 satellite communication systems and, 1434 scattering in, 1271, 1434, 1437-39, 1445 scintillations in, 1436 sleet, snow, hail attenuation in, 1442 subrefraction in, 1435-36 terrain scatter and diffraction in, 1445 time delay in, 1436 transmission paths for, 1443-48 water vapor and absorption in, 1437, 1437 waveguide and, 1434 window regions in, 1433, 1434 Millington method and groundwave propagation, 2060 Min-Idle batching, 235 miniaturization, waveguide, 1405-11, 1406 minicells, 376 minimal encoders, convolutional coding, 600 minimal polynomials, BCH coding, binary, 242-243 minimax criterion equalizers, 81, 82-83 minimum cell rate, ATM, 266, 552 minimum cost tree, ad hoc wireless networks, 2891 minimum distance algorithm, 2056, 2142 minimum functions, BCH coding, binary,, 242 minimum laxity first with time tolerance scheduling, 1555-56 minimum likelihood, quadrature amplitude modulation, 2054 minimum mean square error algorithm (see also Wiener filters), 1116 blind equalizers and, 292 blind multiuser detection and, 298-307 broadband wireless access and, 321 code division multiple access and, 463-464 magnetic recording systems and, 2261 magnetic storage and, 1329 maximum likelihood estimation and, 1341 multiple input/multiple output systems and, 1455 packet rate adaptive mobile receivers and, 1886-1903 minimum mean square error algorithm (see also Wiener filters) (continued) predistortion/compensation in RF power amplifiers and, predistortion in, 532-533 tapped delay line equalizers and, 1690, 1691, 1692 wireless multiuser communications systems and, 1616-17 wireless transceivers, multi-antenna and, 1588 minimum output energy detector, 301 minimum reuse routing, in routing and wavelength assignment in WDM, 2103 minimum shift keying, 584-593, 1457-77 additive white Gaussian noise and, 1468 antipodal signaling in, 1457 bandwidth and, 1457 binary frequency shift keying and, 1457 coherent detection in, 1468-70, 1469 continuous phase frequency shift keying and, 593-598, 1457 continuous phase modulation and, 1457-59, 1458, 2182 continuous shift keying and, 1473-74 cross coupled IQ transmitter for, 1467, 1467 decoder for, 1462-63 differentially coherent detection in, 1470-71, 1471 direct sequency FSK and, 1474 duobinary encoder in, 1474, 1475 frequency shift in, 1457 full and partial response schemes in, 1457, 1458 generalized, 1458 generalized, 1457 inphase-quadrature signal in, 1457, 1463, 1459-61, **1460**, **1461**, 1472 instantaneous frequency pulse in, 1458 intersymbol interference and, 1466-67 maximum a posteriori estimation and, 1472, 1473-74 maximum likelihood sequence estimation (MLSE) and, 1458, 1469-70, 1469 memory and sidelobes in, 1457, 1458 memory transmitter and coherent detection in, 1469-70 memoryless transmitter and coherent detection in, 1468-69 modulation index for, 1457 multiple amplitude MSK and, 1475 normalized phase smoothing response in, 1458 offset quadrature phase shift keying and, 1459-64, 1460, 1461, 1472 orthogonal signaling and, 1457 phase trellis in, 1458-59, 1459, 1467-68, 1468 power control and, 1457 power spectral density of, 1463-64, 1464, 1474 precoded, 1462-63, 1462 receivers and transmitters for, 1462-67, 1462, 1465, 1467 Rimoldi's transmitter for, 1467-68, 1468 serial type, 1464-67, 1465, 1466 serially concatenated coding for CPM and, 2182 sinusoidal frequency shift keying and, 1457, 1458, 1459, 1462, 1473-74 spectral characteristics of, 1463-64, 1464 Sunda's FSK and, 1472 synchronization techniques in, 1471-72, 1471, 1472 traveling wave tube amplifiers, 1457 trellis coded modulation and, 1469-70, 1470 minimum spanning tree, 1532, 1639-40, 2886 minimum variance distortionless response, 1886-1903 mininum mean squared error receiver, 98, 101, 102, 103, 106 - 109minislot usage information, in cable modems, 334-335 min-sum algorithm, low density parity check coding, 1315 miracle octad generator, Golay coding, 886, 888 mirroroptical crossconnects, MEMS, 1705, 1705, 1705 mirrors, lasers, 1776-77 mismatch vector, acoustic echo cancellation, 6, 7 missynchronization detection, Reed-Solomon coding for magnetic recording channels, 471-472 mixed excitation linear prediction, 1266-67, 1300, 1306, 2351, 2351, 2356, 2822-24, 2823 mixed integer programming, routing and wavelength

mixed integer programming, routing and wavelength assignment (RWM) in WDM, 2100

mixers, 139, 1478 MNP compression, 1496 mobile ad hoc network, 2883-99 mobile agents, paging and registration, 1918 mobile application part, global system for mobile, 906 mobile deep range telemetry, 26, 26 mobile IP, 1103, 2938-39, 2938, 2988-89, 2989 mobile nodes, satellite communications, 2118 mobile positioning (see also wireless, location in) mobile radio communications, 308, 1477-84 antenna arrays and, 141 antenna for (see antenna for mobile communications) cdma2000 and, 358-369, 1483 cellular telephony and (see also cellular telephony), 1478-82 channels in, 1481 code division multiple access and, 1481-82, 1482, 1483 control channel in, 1478 diversity techniques in, 1481 Doppler effect, Doppler fading in, 1481 fading in, 1481 frequency division multiple access in, 1481-82, 1482 global system for mobile, 1481, 1482 history of, 1477-78 IMT2000 and, 1094-1108 intersymbol interference and, 1481 multipath interference in, 1481 multiple access techniques for, 1481-82, 1482 orthogonal frequency division multiplexing and, 1867 paging and registration in, 1914-28 RAKE receivers and, 1481 second generation systems in, 1482-83 shadowing in, 1481 simulation and, 2290-91 third-generation systems in, 1483, 1483 time division multiple access in, 1481-82, 1482 trunking theory and, 1478 U.S. digital cellular systems, 1481 in underwater acoustic communications, 46 Walsh coding and, 1482 wideband CDMA in, 1483, 1483 mobile satellite service, 877, 1251, 2112, 2656 mobile station, 376, 380, 905-17 mobile station antenna, cellular telephone, 192-196 Mobile Station Base Station Compatibility Standard for Dual Mode Wideband...Cellular, 347 mobile switching center, 1479 mobile telephone ISDN number, 906 mobile telephone switching office, 1602 mobility indexes, in paging and registration, 1917, 1919 mobility management general packet radio service and, 869-870 global system for mobile and, 914-915 satellite communications and, 2116-17, 2118, 2119-20 wireless packet data and, 2987-89, 2987 mobility portals and services, 2190-96 addressing, IP addressing, 2194 benefits of, 2192 billing systems for, 2194 compact HTML and, 2193 composite capabilities/preference profiles and, 2194 differentiated services and, 2195 general packet radio service and, 2191, 2193 global system for mobile and, 2192, 2193 iMode and, 2193 integrated services and, 2195 Internet and, 2190-91 Internet relay chat and, 2192 interoperability issues and, 2195 Jini and, 2194 multiprotocol label switching and, 2195 personal digital assistants and, 2190, 2191, 2194 quality of service, 2192, 2195 resource reservation protocol and, 2195 rise of, 2191 Salutation and, 2194 second generation wireless systems and, 2192 security and, 2194-95 short message service and, 2190

smartphones and, 2191 terminals and, 2193-94, 2194 universal mobile telecommunication service and, 2191, 2194 Universal PnP and, 2194 videoconferencing and, 2195 VoiceXML and, 2192 wireless application protocol and, 2190, 2191, 2192-95, 2193 wireless identity module and, 2195 modal dispersion, in optical fiber, 1507 modal noise, in optical fiber systems, 1843 mode converting (long period) gratings, 1723, 1723, 1727-28, 1727, 1728 mode locked lasers, 1762, 1961 model-based admission control, 117 modeling channel (see channel modeling and estimation) chaotic systems and, 422 traffic (see traffic modeling) modeling and analysis of digital optical communication systems (see also optical fiber), 1484-94 additive white Gaussian noise in, 1487 amplifiers in, 1484, 1485, 1486 Brillouin scattering and, 1491 cross phase modulation and, 1490-91, 1490 development of, 1484 differential group delay in, 1492-93 dispersion compensating module in, 1484 distortion in, 1484-85 erbium doped fiber amplifier in, 1484 four-wave mixing in, 1490 Gordon-Haus effect in, 1490 interference in, 1484 Jones vectors and Jones matrix in, 1492 multiplexing/demultiplexing in, 1484 noise and, 1484-85, 1486 noise figure in, 1485 nonlinear Schrodinger equation in, 1488, 1489, 1491 optical birefringence in, 1492 polarization and, 1484-85, 1491-93 polarization dependent loss in, 1493 polarization mode dispersion in, 1492-93 principal states of polarization in, 1492 Raman scattering in, 1491 receivers (coherent receivers) for, 1484, 1486-88, 1486, 1487, 1488 scattering in, 1491 self-phase modulation and, 1489, 1489 signal to noise ratio in, 1485-87, 1493 spectral efficiency of, 1488 transmitters for, 1484, 1488-91, 1489, 1491 wavelength division multiplexing and, 1490-91, 1490, 1484-85, 1485 modems, 539, 1494-1501 acoustic telemetry in, 23-24 amplitude modulation and, 1497 analog to digital conversion in, 1495 asynchronous mode in, 1495 baud in, 1496 bit rate in, 1496 cable (see cable modems) compression in, 1496 cyclic redundancy check in, 1495 data communication equipment and, 1495, 1496 data terminal equipment and, 1495, 1496 differentail PSK and, 1497 digital subscriber line and, 1499-1500 digital to analog conversion in, 1495 DOCSIS and, 1500 fax transmission using, 1499 file transfer/data transfer in, 1497 flow control in, 1496-97 forward error correction in, 1497 frequency modulation and, 1497 frequency shift keying and, 1497, 1498 H.324 standard and, 919-920 high frequency communications and, 951, 952-955 home area networks and, 2687 integrated services digital network and, 1495 Internet service providers and, 1498-99

3038 INDEX

modems (continued) link access protocol-modem protocol in, 1496 MNP compression in, 1496 modulation techniques in, 1497-99 Naval Undersea Warfare Center range-based modem in, 25-26, 25 operation of, 1495-97 parallel vs. serial transmission and, 1494-95 phase shift keying and, 1497 POTS splitters for, 1500 public switched telephone network and, 1495 pulse amplitude modulation and, 2022 pulse coding modulation and, 1497 quadrature amplitude modulation and, 1497, 1498 retraining in, 1498 RS-232 connections in, 1495 RTS/CTS flow control in, 1497 shell mapping and, 2222, 2227 software radio and, 2308 speeds of, 1495-96 standards for, 1497-99 start/stop transmission in, 1495 symbols in communication and, 1497 synchronous mode in, 1495 telesonar, 2215 trellis coded modulation and, 1497, 1498, 2632, 2632, 2633 two dimensional encoding, 1497 underwater communications (see acoustic modems, underwater communications) universal asynchronous receiver/transmitter and, 1495 V.34, 2640 V.42bis compression in, 1496 V.90, 2770-79, 2771 very high speed DSL and, 2779 Vxx standards for, 1498-99 X2 protocol for, 1498, 1499 Xmodem, Ymodem, Zmodem in, 1497 Xon/Xoff mechanisms in, 1496 modenumber, modespacing, lasers, 1777 modification attacks, 1151 modified frequency modulation, 576, 1327 modified prime coding, optical synchronous CDMA systems, 1812, 1819–21 modified syndromes for decoding, in BCH and Reed-Solomon coding, 259-260, 469-470, 469 MODTRAN database and calculations, free space optics, 1856-57 modulation, 16, 132-141, 371, 1477-78, 2179 in acoustic modems for underwater communications, 19 acoustic telemetry in, 23, 24 active antenna and, 49 amplitude (see amplitude modulation) bit interleaved coded modulation and, 275-286, 275, 277 bit interval in, 7, 1335 carrierless amplitude and phase, 336-339, 337, 338 cdma2000 and, 362 CDROM and, 1735 chaotic systems and, 422, 423, 423 chirp, 440-448 continuous phase coded, 584-593 continuous phase frequency shift keying and, 593-598 continuous phase modulation and, 593-598 demodulation process and (see also demodulation), 1335 digital audio broadcasting and, 684 digital phase modulation and, 709-719 duobinary frequency shift keying, 585 free space optics and, 1851 frequency modulation in, 807-825 frequency shift keying and, 593 Gaussian minimum shift keying and, 584-593 generalized tamed frequency modulation and, 585 global system for mobile and, 913 high frequency communications and, 954 IS95 cellular telephone standard and, 350, 351, 353, 354 lasers and, 1779, 1781 leaky wave antenna and, 1241

magnetic recording systems and, 2249-51, 2253-58 M-ary frequency shift keying and, 1335 M-ary time shift keying and, 1335 matched filter demodulation and, 1335-38, 1336 minimum shift keying and, 584-598, 1457-77 modems and, 1497-99 multiple input/multiple output systems and, 1455-56 optical fiber and, 1708, 1825–32, 1848 optical modulators and, 1741–48 optical synchronous CDMA systems and, 1809, 1813 partial response signals and, 1928, 1933-34 permutation coding and, 1953 phase modulation in, 807-825 phase shift keying and, 593, 1335 power spectra of digitally modulated signals, 1988-95 powerline communications and, 1995, 2003 predistortion/compensation in RF power amplifiers and, 530 pulse amplitude modulation and, 1335, 2021-30 pulse position modulation, 2030-42, 2031 quadrature amplitude modulation, 1335, 2043-58, 2043 quadrature phase shift keying and, 589 raised cosine modulation and, 585 satellite communications and, 1225, 1225 satellite onboard processing and, de- and remodula-tion in, 480–482, **480**, **481** spectrally raised cosine modulation and, 585 speech coding/synthesis and, 2368 symbols and symbol duration in, 1335 tamed frequency modulation and, 584-593 terrestrial digital TV and, 2549-50 trellis coded modulation and, 590 trellis coding and, 2635 in underwater acoustic communications, 40-41, 45-46 wideband CDMA and, 2878 wireless infrared communications and, 2927-27 wireless multiuser communications systems and, 1610, 1611-12 modulation coding (MC), 570, 573, 1326 modulation index, in minimum shift keying, 1457 modulation transfer function, constrained coding techniques for data storage, 572-573, 572 MOE algorithm, adaptive receivers for spread-spectrum system, 109 Molnya orbit, 1250 moment methods, 403, 406-407 monitoring signal quality (see signal quality monitoring) monofractal models, traffic modeling, 1669 monopole antenna, 142, 193, 193, 193 monotonic transformation, maximum likelihood estimation, 1339 Monte Carlo simulation, 422, 2285, 2291-94, 2916 Moore's law, 263 morphological operators, in image processing, 1074 motion estimation and compensation, 1042-44, 1042 Motley-Keenan model, indoor propagation models, 2016. 2016 moving average filters, in channel modeling, estimation, tracking, 412, 413 moving coil electrodynamic loudspeaker transducers (acoustic), 34, 34 MP3 audio compression, 1319 MPEG compression, 232 BISDN and, 263 cable modems and, 324, 330 community antenna TV and, 522, 525 digital audio broadcasting and, 682-683 digital versatile disc and, 1738 image and video coding and, 1029, 1052, 1053-55, 1054 magnetic storage and, 1319 medium access control and, 1558, 1559 multimedia over digital subscriber line and, 1571 piggybacking in, 232-233, 232 satellite communications and, 880 speech coding/synthesis and, 2356, 2819 streaming video and, 2435-36 terrestrial digital TV and, 2552-53, 2553 traffic modeling and, 1672 video, unequal error protection coding and, 2765-66

vocoders and, 2819 wireless MPEG 4 videocommunications and, 2972-81 MSAT satellite communication, 196, 198 Muller matrix, optical fiber, 436 multiaccess interference, 2215 multiaddress set claim protocol, 1536-37, 1536 multiband excitation coding, speech coding/synthesis, 2351 multibase ALOHA, 131 multibeam phased arrays, 1513-21, 1514 analog to digital conversion and, 1520, 1521 applications for, 1519 bandwidth and, 1518 beam steering using, 1519, 1520 beamforming in, 1517, 1517, 1518, 1520-21, 1520 beamwidth in, 1517 Butler beamformer using, 1517, 1518 coding division multiplexing in, 1514 cost of, 1518 element arrangement in, 1513-14 far field in, 1514 Fraunhofer region and, 1514 frequency reuse and, 1514 Fresnel region in, 1514 future developments for, 1521 gain in, 1517 interference cancellation in, 1519, 1520-21 Iridium and, 1519 linear arrays in, single and multibeam, 1514-15, 1515, 1516 mutual coupling in, 1516 near field in, 1514 one dimensional, 1513 overlap of beams in, 1516 polarization in, 1514 satellite communications and, 1519 scan loss in, 1517 sensitivity and, 1518-19 sidelobe level in, 1517 signal to noise ratio and, 1519 simultaneous users and, transmit and receive limits, 1517 time division multiplexing in, 1514 tracking and data relay satellite using, 1519 two dimensional, 1513, 1515-17, 1516, 1517 multicarrier CDMA, 1521-28 additive white Gaussian noise and, 1522, 1526 comparison of various flavors of, 1527-28 direct sequence CDMA and, 1522, 1523 direct sequence spread spectrum and, 1521, 1523 equal gain combining detector in, 1527 fast Fourier transforms in, 1522 frequency division multiplexing and, 1522 frequency domain coding in, 1524 interference in, 1527 maximal ratio combining detector in, 1527 maximum likelihood detector in, 1527 multicarrier direct sequence CDMA, 1524 multicarrier modulation in, 1521-28, 1522 multitone CDMA and, 1525 multiuser interference and, 1527 multiuser OFDM and, 1527-28 orthogonal frequency division multiplexing in, 1521, 1523, 1524, 1525–28, **1526** orthogonality restoring detector in, 1527 RAKE receivers and, 1523-24, 1523 spreading sequence in, 1523 transmitter and receiver for, 1522-25, 1522, 1524, 1525 Walsh-Hadamard coding in, 1526 multicarrier direct sequence CDMA, 1524 multicarrier frequency division duplex, 2116 multicarrier modulation asymmetric DSL and, 1572 multicarrier CDMA and, 1521-28, 1522 multicarrier transmission, synchronization, 2481-82 multicast algorithms, 1529-38 multicast backbone, 1535, 2432 multicast capable OXCs, 2104 multicast open shortest path first, 1533-34, 1534

multicast routing, wavelength assignment, 2104-05 multicast source discovery protocol, 1535 multicast tree, 1530, **1530** multicasting, 2615 ad hoc wireless networks and, 2890-93 border gateway multicast protocol and, 1535, 1536 border gateway protocol and, 1535 core-based tree in, 1535 distance vector multicast routing protocol in, 1534 Ethernet and, 1529-30 forwarding in, 1532 group communication and, 1529-31, 1529, 1530 interdomain routing protocols in, 1535-37 Internet and, 1531-32, 1531 intradomain routing and, 1533-35 IP addresses in, 1531 IP networks and, 1531 KMB algorithm in, 1533 Kruskal's algorithm in, 1532 local area networks and, 1529-32 MAC addresses and, 1529, 1531-32, 1532 metropolitan area networks and, 1529 minimum spanning tree problem in, 1532 multiaddress set claim protocol and, 1536-37, 1536 multicast backbone and, 1535 multicast open shortest path first in, 1533-34, 1534 multicast source discovery protocol in, 1535 multicast tree in, 1530, 1530 multiprotocol extension to BGP4, 1535 parent and child targets in, 1536 point to point networks and, 1530-31, 1530 Prim's algorithm in, 1532 protocol independent multicast sparse mode in, 1534-35 reachability in, 1536 reverse shortest path tree and, 1534 routers and, 1531, 1531, 1533 routing information protocol and, 1534 security and, 1154-55 shared-medium networks and, 1529-30 Steiner tree problem in, 1532-33 target lists in, 1536 TM algorithm and, 1533 token ring and, 1529 unicast vs., 1530 wavelength division multiplexing and, 655 wide area networks and, 1532 multicasting core extraction distributed ad hoc routing, 2892-93 multicavity optical filters, 1725 multichannel ALOHA, 130 multichannel multipoint distribution services, 2655, 2671 multicopy ALOHA, 130 multidimensional coding, 1538-51 additive white Gaussian noise and, 1542-43, 1542, 1545-47, 1546 algebraic replicas in, 1541-42 bit error rate, 1545–48, **1548** burst errors and, 1540–41, 1544–48 bursty channels and, 1547-48 cdma2000 and, 1548 concatenated MDPC coding, 1548–49, 1549 density in SPC coding, 1540 error control coding and, 1539-40 error detection and correction in, 1540-41, 1544-48 Galois fields and, 1538, 1542 Gilbert coding as, 1540 Hamming coding and, 1541 hard vs. soft decision decoding in, 1541 Hobbs coding as, 1540 iterative coding and, 1538, 1543, 1544 loglikelihood ratio in, 1541-42 maximum a posteriori (MAP) algorithm and, 1542 palindromic coding and, 1540 parity check coding in, 1543-44 product coding as, 1538-40, 1539 reciprocal coding in, 1540 rectangular parity check coding in, 1543-44 replica in SPC coding, 1541-42 single parity check coding, products of, 1540-43

trellis coding and, 1538 turbo coding and, 1548-49, 1549 two dimensional burst coding as, 1538 two dimensional dot coding as, 1538 wideband CDMA, 1548 multifractal models, in traffic modeling, 1670 multihop satellite routing, 1254 multihop WDM networks, 1551 multilayer perceptrons, in neural networks, 1678 multilevel or polybinary signals, in optical receivers, 1825 multimedia adaptive round robin and earliest available time scheduling in, 1559 digital subscriber line and, 1570-79 hybrid optical networks and, 1559 medium access control and, 1558-60, 1559 MPEG compression and, 1559 session initiation protocol and, 2196-2206 multimedia cable network system, 272, 324 multimedia MAC protocols for WDM optical networks (see also medium access control), 1551-62 multimedia messaging service center, 2973 multimedia messaging services, 2978-79, 2979 multimedia mobile access communication, 2941 multimedia networking (see also multimedia), 1562-69 asynchronous transfer mode and, 1567 bandwidth and, 1562, 1563, 1565, 1568 bandwidth brokers and, 1568 broadband integrated services digital network and, 1567 buffer management in, 1562, 1563, 1565-66 call admission control in, 1563-64 congestion control and, 1566 constant bit rate in, 1563, 1566, 1567 constraint-based routing using label distribution protocol, 1568 deficit round robin in, 1565 delay in, 1562 differential services coding point and, 1568 differentiated services and, 1568-69 first come first served in, 1565 first in first out algorithm for, 1564, 1565 forwarding in, 1568 functional requirements of, 1562 generalized processor sharing scheduler in, 1565 history and development of, 1567 integrated services digital network and, 1563 Internet and, 1567-69 Internet integrated services architecture for, 1567 isolation in, 1564 jitter in, 1562 label distribution protocol in, 1568 linearly bounded arrival process in, 1568 macroflows and, 1568 multicast routing and, 1566 multiprotocol label switching and, 1568 overprovisioning in, 1563, 1567 packet dropping in, 1565-66 per hop behavior in, 1568 quality of service and, 1562-68 queue partitioning in, 1565 reliability in, 1562 reservation protocols and, 1567 resource allocation for, 1563, 1567 resource reservation protocol and, 1567, 1568 routing, for QoS in, 1563, 1566 scaling mechanisms for, 1563, 1566 scheduling in, 1563, 1564-65 skew in, 1562 traffic requirements of, 1562 traffic shaping algorithms in, 1563, 1564 traffic verification or policing algorithms in, 1563 transmission control protocol and, 1566 variable bit rate in, 1563, 1566, 1567 virtual private networks and, 1568 weighted fair queue in, 1564, 1565 weighted round robin queuing in, 1564, 1565 multimedia over digital subscriber line, 1570-79 applications for, 1576-77 architecture design for, 1572, 1573

asymmetric DSL and, 1570, 1571-72, 1571 automatic repeat request in, 1571 bit error rate and, 1573-75 channel gain to noise ratio in, 1573 compression and, 1570-71 cost minimization in, 1573-75 error resilient entropy coding in, 1576 image data over, 1576 integrated services digital networks and, 1570 joint source and channel optimization in, 1571 layered coding in, 1570-71 MPEG compression and, 1571 multicarrier modulation in, 1572 parallel and serial transmission in, 1572, 1574-75 peak signal to noise ratio in, 1576-77, 1576, 1577 quadrature amplitude modulation and, 1576 quality of service and, 1571, 1573, 1575-76 signal to noise ratio in, 1573 subchannel to layer assignment in, 1574-75 system optimization in, 1572-76 time slot assignment in, 1574 video over, 1576 multimedia wavelength division multiple access, 1558 multimode coding, in speech coding/synthesis, 2354-55 multimode interference coupler, 1761 multimode optical fiber, 434, 1507, 1707, 1842 multimodulus algorithm, blind equalizers, 292 multipass optical filters, 1725 multipath (see also fading; interference), 781-802, 2065, 2067 multipath fading in acoustic modems for underwater communications, 15 acoustic telemetry in, 22 adaptive receivers for spread-spectrum system and, 95, 105-108 antenna for mobile communications and, 190 bit interleaved coded modulation and, 276, 278 cable modems and, 334 cellular communications channels and, 393, 394 chann/in channel modeling, estimation, tracking, 398, 410 digital audio broadcasting and, 677, 678, 685-685, 685 diversity and, 730-731 indoor propagation models and, 2013, 2013, 2018 intelligent transportation systems and, 509-510 local multipoint distribution services and, 1273, 1277 location in wireless systems and, 2967-68 microwave and, 2562-65, 2562 millimeter wave propagation and, 1434, 1438, 1445 mobile radio communications and, 1481 packet rate adaptive mobile receivers and, 1886 polyphase sequences and, 1975 power control and, 1983 satellite communications and, 196, 197, 1226-27, 1226 shallow water acoustic networks and, 2207 simulation and, 2290-91 space-time coding and, 2324 spatiotemporal signal processing and, 2333 tropospheric scatter communications and, 2697-98 ultrawideband radio and, 2761 underw/in underwater acoustic communications, 38 Viterbi algorithm and, 2817-18, 2817 wireless and, 2916-18 wireless multiuser communications systems and, 1603. 1603 multipath channels and signals, blind multiuser detection, 302-306 multipath delay spread, wireless multiuser communications systems, 1604 multipath diversity, IS95 cellular telephone standard, 355-356 multiple access and IS95 cellular telephone standard, 354 multiple access channels and protocols, in point to point communications, 339 multiple access collision avoidance, 1347-48, 2210, 2212, 2885 multiple access collision avoidance wireless, 2210, 2885-86

multiple access interference adaptive receivers for spread-spectrum system and, 97-98, 101-102, 103 blind multiuser detection and, 299, 303 chirp modulation and, 445, 446 code division multiple access, 458-446, 1196, 2278, 2283 optical fiber and, 1809 optical synchronous CDMA systems and, 1809, 1810 packet rate adaptive mobile receivers and, 1886, 1887 polyphase sequences and, 1976 signature sequence for CDMA and, 2278, 2283 spatiotemporal signal processing and, 2333, 2336 synchronization and, 2479-85 wireless multiuser communications systems and, 1615 multiple access systems, adaptive receivers for spreadspectrum system, 95-96, 96 multiple amplitude MSK, 1475 multiple antenna transceivers for wireless communications, 1579-90, 1580 additive white Gaussian noise and, 1580 Alamouti scheme in, 1584-85 average capacity criterion in, 1589 beamshaping in, 1579 block space time multiplexing in, 1589 closed loop capacity in, 1581, 1582, 1583, 1585 combined transmit and receive diversity systems for, 1586-87, 1587 correlated vs. uncorrelated antenna elements in, 1583 diversity vs. power gain in, 1583, 1584 fading in, 1579 gain and, 1579, 1583, 1584 interference in, 1579 minimum mean squared error and, 1588 multiple input/multiple output systems and, 1580, 1581 - 82multiple input/single output systems in, 1582-83 multiplexing/demultiplexing in, 1580 open loop capacity in, 1581-83, 1584 orthogonal frequency division multiplexing in, 1582 random capacity concept in, 1580-81 receiver options for capacity attainment in, 1583 scattering in, 1579 Shannon or channel capacity, 1579, 1580-82 signal to interference plus noise ratio in, 1579 signal to noise ratio and, 1579, 1580, 1582, 1583, 1584 single input/multiple output systems in, 1582-83 single receiver antenna systems, 1584 smart antenna systems and, 1580 space time transmission schemes for, 1584-85 spectral efficiency in, 1579, 1581-82 substreams in, 1580 switch transmit diversity in, 1586 transmit MRC in, 1585-86 vertical Bell Labs layered space time scheme in, 1587-89. 1589 zero force projection in, 1588 multiple frequency shift keying, 16, 19, 23, 24 multiple increase and linear decrease, 2886 multiple input multiple output systems, 1450–56, 1450 array gain in, 1450, 1450 block coding and, 1455 in channel modeling, estimation, tracking, 400-401, 400 channel model for, 1452-53, 1452 cochannel interference and, 455 code division multiple access and code division multiple access, 1455 direct sequence CDMA in, 1456 diversity gain in, 1450-52, 1451, 1452 diversity in, 1455 Doppler effect, Doppler spread in, 1453 encoding in, 1455-56 fading, Rayleigh fading in, 1453–55, 1454 gain in, 1450–1453 interference and, 1119 interference reduction and, 1450, 1452, 1452 intersymbol interference and, 1455 link adaptation in, 1455

maximum likelihood detection in, 1455 maximum likelihood sequence estimation and, 1455 minimum mean squared error and, 1455 modulation for, 1455-56multiple input/single output systems and, 1451 multiplexing gain and, 1450, 1452, 1452 orthogonal frequency division multiplexing in, 1456, 1878 parity coding and, 1456 receivers in, 1450, 1455 Reed-Solomon coding and, 1456 Shannon or channel capacity in, 1453-55, 1454 signal to interference ratio in, 1452 signal to noise ratio in, 1453-54 signaling in, 1455-56 single input/multiple output systems and, 1451 single input/single output systems and, 1451 space-time coding and, 2327, 2330 spatial multiplexing and, 1452, 1455 training mode in, 1453 transmitters in, 1450 trellis coding and, 1455 turbo coding and, 1456 very high speed DSL and, 2803-05 Viterbi algorithm/decoder in, 1455 wireless transceivers, multi-antenna and, 1580, 1581-82 zero forcing receivers and, 1455 multiple input/single output systems multiple input/multiple output systems and, 1451 wireless transceivers, multi-antenna and 1582-83 multiple instruction/multiple datastream, 2313 multiple link admission control, 117-118 multiple link service access points, 1281 multiple quadrature amplitude modulation, 24 multiple quantum well modulators, 1967, 1968 multiple sample and hold converter, 2234 multiple stack algorith, in sequential decoding of convolutional coding, 2159-60 multiplexers/multiplexing adaptive receivers for spread-spectrum system and, <u>9</u>5 cdma2000 and, 359, 363 H.324 standard and, 920-922, 921 IS95 cellular telephone standard and, 350, 352 media access control and, 1342 optical communications systems and, 1484 optical, 1748-59, 1748 orthogonal transmultiplexers and, 1880-85 packet switched networks and, 1907-09 satellite onboard processing and, 482 statistical, 2420-32 synchronous digital hierarchy and, 2498, 2498, 2499, 2500, 2500-04, 2501-2506 transmission control protocol and, 2604 transport protocols for optical networks and, 2617-18 wireless transceivers, multi-antenna and, 1580 multiplexing gain, in MIMO systems, 1450, 1452, 1452 multipoint communication, multicasting, 1529-31, 1529, 1530 multipoint control unit, IP telephony, 1174 multipoint to multipoint communications, adaptive receivers for spread-spectrum system, 95 multiprotocol extension to BGP4, 1535 multiprotocol label switching, 549, 1590-1601, 2654 admission control and, 116 applications of, 1599-1600 architecture of, 1593-99 asynchronous transfer mode and, 1594, 1598-99 bandwidth allocation in, 1598 border gateway protocol 4 (BGP4) and, 1597 broadband and, 2655, 2674-75, 2674 congestion control in, 1594, 1599 constraint-based label distribution protocol in, 1596 control in, independent vs. ordered, 1595 differentiated services and, 1594, 1597, 1598 distribution in, unsolicited vs. on-demand, 1595 downstream vs. upstream allocation in, 1595 early congestion notification and, 1594 equal cost multipath in, 1599 explict routed LSP in, 1592, 1593, 1593

fault tolerance and, 1640 FEC to NHLFE in, 1594 flow control, traffic management and, 1658, 1658, 1659 forward equivalence class in, 1591 forwarding in, 1591, 1593-95 hierarchical forwarding in, 271 history and development of, 1598-99 hop by hop routed LSP in, 1591–92, **1593** implementation of, 1599 incoming label map table for, 1591, 1594 integrated services and, 1597 Internet protocol and, 1590-1601 IP networks and, 271, 1590-1601 label distribution protocol in, 1591, 1593, 1595-96 label edge routers in, 1591 label encapsulation in, 1594, 1594, 1597-98 label merging in, 1595 label operations, label stacks in, 1594 label spaces in, 1595 label switched patch in, 271, 1590, 1591-93, 1593 label switched routers in, 271, 1591, 1594 label use method in, 1595 labels in, 1591 layer 2.5 architecture and, 1594 link layer specific encapsulation in, 1594 link layer technologies and, 1594 mobility portals and, 2195 multimedia networks and, 1568 next hop label forwarding entries in, 1591, 1592, 1593-94 optical cross connects/switches and, 1798 optical fiber and, 2615 packet switched networks and, 1909 penultimate hop popping in, 1595 protection switching in, 1600 quality of service and, 1597-98 reliability and, 1640 resilience in, 1600 resource reservation protocol for tunneling and, 1596-97 retention in, liberal vs. conservative, 1595 shim headers in, 1594 signaling layer in, 1595-97 statistical multiplexing and, 2420-32 time to live field in, 1594-95 traffic engineering and, 271, 1599 tunneling in, 271 virtual circuit emulation in, 271 virtual private networks and, 1591, 1599-1600, 1600, 2809, **2809** multiprotocol over ATM, 1599 multipulse LPC, in speech coding/synthesis, 2348, 2355 multirate services, power control, 1987 multiresolution analysis, wavelets, 2851-56 multiscope routing protocols, in ad hoc wireless networks, 2889 multistage detector, in code division multiple access, 464, 464 multistage filters, in packet rate adaptive mobile receivers, orthogonal, 1892, 1892, 1893 multistage interconnection networks, ATM, 202, 202 multistage Wiener filters, adaptive receivers for spreadspectrum system, 104 multistep linear prediction, blind equalizers, 295-296 multistrip couplers, surface acoustic wave filters, 2452-54, **2453** multitone CDMA, 1525 multitone interference, 1130-41 multitrack coding, in constrained coding techniques for data storage, 582 multiuser channel estimation, 771, 1614-15 multiuser communication systems, code division multiple access, 461 multiuser detection cochannel interference and, 455 code division multiple access and, interference cancellation in, 462-465 neural networks and, CDMA, 1680-81 power control and, 1987-88 multiuser interference, multicarrier CDMA, 1527

multiuser OFDM, multicarrier CDMA, 1527-28 multiuser wireless communication systems (see also cellular telephony; wireless), 1601-24 additive white Gaussian noise in, 1605-07 Alamouti scheme in, 1584-85, 1611, 1611, 1619 analog to digital conversion in, 1609 automatic repeat request in, 1612 bandwidth in, 1603, 1604 base station location in, 1603 base station receivers for, 1612, 1613 binary phase shift keying and, 1610, 1614 block fading channels in, 1605 Bluetooth and, 1602 cdma2000 and, 1602 cells in, 1602 cellular telephony and, 1602-03, 1602 channel coding in, 1604 channel coherence bandwidth in, 1604 channel coherence time in, 1604 channel decoding in, 1609, 1618-19 cochannel interference in, 1604 code division multiple access and, 1602, 1608, 1609, 1615 constant modulus algorithms in, 1614 convolutional coding and, 1609-10, 1609, 1610 decorrelating detectors in, 1616 delay power spectrum of channel in, 1604 digital to analog conversion in, 1609 direct sequence CDMA and, 1602, 1608, 1614, 1615 diversity in, 1603, 1608 Doppler power spectrum in, 1604 Doppler spreading in, 1604 ergodicity in, 1606 fading in, 1603-11 filters in, 1616 flat fading in, 1604 forward error correction in, 1609 frequency diversity in, 1603 frequency division multiple access in, 1602 frequency nonselection (flat fading) in, 1604 frequency reuse in, 1608 frequency selective channels in, 1605 generator polynomials, in encoding, 1610 handoffs in, 1602 interference cancellation in, parallel, 1617-18 interference cancellation in, subtractive and successive, 1617 interference in, 1604 intersymbol interference in, 1612, 1616 layered architecture in, 1602-03 linear detectors in, 1616-17 local area networks and, 1602 low density parity check coding in, 1610 matched filter detectors in, 1616 maximal ratio combining in, 1619 maximum a posteriori detectors in, 1616 minimum mean squared error detectors in, 1616-17 mobile telephone switching office in, 1602 modulation in, 1610, 1611-12 multipath delay spread in, 1604 multipath interference in, 1603, 1603 multiple access interference and, 1615 multiuser channel estimation in, 1614-15 multiuser detection in, 1615-18 normalized sum rate in, 1608 personal area networks and, 1602 power control in, 1606, 1619 power management in, 1604 quadrature phase shift keying and, 1610 received signal characteristics in, 1613-14 reliability in, 1605 sectorization of cells in, 1603 Shannon or channel capacity in, 1605-08, 1606 shared multiple access in, 1603-04 signal to noise ratio in, 1606, 1619 spatial diversity in, 1603 spreading in, 1611-12, 1612 third-generation systems in, 1602 time division multiple access and, 1602, 1609 time varying multipath in, 1603, 1603 traffic activity factors in, 1608

training symbols in, 1614-15 transceiver architecture for, 1608-20 trellis coded modulation and, 1610 trellis diagrams in decoding, 1618, 1618 turbo coding in, 1604, 1610 underspreading in, 1604 Viterbi algorithm in, 1619 water filling in time in, 1606 wide area networks and, 1602 wideband CDMA in, 1602, 1608 wireless LAN and, 1602 multiwavelength optical network, 1509 mutual coupling adaptive antenna arrays and, 73-77, 74, 75, 76, 77 microstrip/microstrip patch antenna and, 1370-71, 1371, 1372, 1373 multibeam phased arrays and, 1516 mutual information rate, rate distortion theory and, 2070 Nagle algorithm, transmission control protocol, 2609 NAHJ algorithm, blind multiuser detection, 301, 302, 306 Nahuel horn antenna, waveguide, 1392, 1392 Nakagami fading, 785–786 diversity and, 733 power control and, 1983 quadrature amplitude modulation and, 2050-52 wireless and, 2919 Nakagami m distribution, cellular communications channels, 394 naming conventions, 547-549 narrow band digital broadcasting, 680 narrow sense BCH coding, 622 narrowband communications in channel modeling, estimation, tracking, 409–410 indoor propagation models and, 2015–17, **2015** microstrip/microstrip patch antenna and, 1357 powerline communications and, 1996-97 sampling and, 2108-11, 2109 space-time coding and, 2324 speech coding/synthesis and, 2341 synchronization and, 2473-85 narrowband interference, 1130-41, 2002 National Science Foundation, 268 Naval Undersea Warfare Center range-based modem, 25-26. 25 Navigation System with Time and Ranging, 198 Nd/YAG lasers, free space optics, 1853 near end crosstalk VDSL, 2786, 2798-2800, 2804-05 near far effect, power control, 1982-83, 1983 near field, loop antenna, 1292, 1293, 1514 near grazing incidence, radiowave propagation, 210 nearest neighbor condition in quantization, 2129 nearest neighbor problem, 642, 2126 near far problem, 1343, 1680, 2208 adaptive receivers for spread-spectrum system and, code division multiple access and, 458, 461-462 negaperiodic complementary sequences, 898-899 negative acknowledgment, 226, 2211 Net audio, 544 NETBLT, 2616 Netto's constructions, in low density parity check coding, 661 network access points, IP networks, 268 network address translation, 1651 network allocation vector, wireless LAN, 1287 network coverage, 377, 378 network flow control (see flow control, network) network identifier, 269, 548 network information theory, 1114-15 network interface layer, TCP/IP model, 541 network layer, OSI, 15 OSI reference model, 539 packet switched networks and, 1911 shallow water acoustic networks and, 2217-18, 2218 TCP/IP model, 541 network layer security, 1153-54 network network interface, 113, 264, 265 network reliability and fault tolerance (see also fault tolerance; reliability), 1631-44

network security (see security) network traffic management (see traffic management) network traffic modeling (see traffic modeling) neural networks, 1675-83 activation function in, sigmoidal, 1676 admission control and, 1681 antipodal, 1676 applications of, in communications, 1679–81 architectures for, 1676–77 artificial, 1675 associative memory in, 1677 asynchronous transfer mode and, 1681 automatic speech recognition and, 2378-79 back propagation algorithm in, 1678 bias in, 1676 channel modeling and identification using, 1679 competitive learning in, 1678 delta rule in, 1677-78 energy function in, 1677 equalization using, 1679-80 feedforward vs. feedback systems in, 1676-77 generalization in, 1675, 1677-79 hidden layers in, 1676, 1676 Hopfield, 1677 induced local fields in, 1676 learning in, 1675, 1677-79, 1677 local gradients in, 1678 multilayer perceptrons in, 1678 multiuser detection using, 1680-81 network applications for, 1681 neuron nodes in, 1675-76, 1676 pattern classification/association in, 1677 perceptrons in, 1676, 1678 photonic analog to digital conversion and, 1966-68, 1967 processing elements in, 1676 quality of service and, 1681 radial basis functions in, 1678 recall in, 1675 recurrent, 1680 self-organizing map in, 1678 self-organization in, 1678 speech coding/synthesis and, 2378-79 synapses in, 1675–76, **1676** training in, 1675, 1677–79, **1677** transfer function in, 1676 universal approximators using, 1679 unsupervised learning in, 1678 weighting in, synaptic weight, 1676 Window-Hoff rule in, 1677-78 neurocomputing, 1675 neuron nodes, in neural networks, 1675-76, 1676 neurotechnology, 1681 Newton's identities in BCH (nonbinary) and Reed-Solomon coding, 255, 257, 623 Newtonian physics, chaos, 421 Newtonian telescope, free space optics, 1863, 1864 next hop label forwarding entries, 1591, 1592-94 next hop routing, 549 node cover, in optical Internet, 2468 nodes, in sequential decoding of convolutional coding, 2142 noise (see also interference; multipath; signal to noise ratio), 2067 acoustic echo cancellation and, 4, 5 adaptive antenna arrays and, 71-72, 74, 74 adaptive equalizers and, 79 amplitude modulation and, 134 angle modulation methods and, 815-823 atmospheric, 2405-12 automatic speech recognition and, 2378 cable modems and, 331 chaotic systems and, 421-422 community antenna TV and, 512, 514-517, 515, 523-524 constrained coding techniques for data storage and, 573 digital magnetic recording channel and, 1325 free space optics and, 1857-59 frequency division multiple access and, 826-827 frequency synthesizers and, 836, 842-843

noise (see also interference; multipath; signal to noise ratio) (continued) high frequency communications and, 949 impulsive, 2402-2420 magnetic recording systems and, 2257 optical communications systems and, 1484-85, 1486 optical fiber and, 1824-40, 1843-48 optical signal regeneration and, 1759, 1760 orthogonal frequency division multiplexing and, 1874-76 packet rate adaptive mobile receivers and, 1886 parabolic and reflector antenna and, 1922, 1926-27 partial response signals and, 1928 photodetectors and, 996-997 powerline communications and, 2001-2002, 2002 radiowave propagation and, 2061 satellite communications and, 1224-25 sequential decoding of convolutional coding and, 2144 sigma delta converters and, 2229, 2231-32, 2231, 2232 speech coding/synthesis and, 2353-54 turbo trellis coded modulation and, 2738 underw/in underwater acoustic communications, 37 very high speed DSL and, 2789-90, 2790 wireless and, 2915 wireless infrared communications and, 2926-27 noise equivalent power free space optics and, 1858-60, 1859 optical transceivers and, 1833 photodetectors and, 997 noise figure, in optical communications systems, 1485 noise predictive maximum likelihood, 2248, 2250, 2261-66, 2261, 2262 noise subspace, blind multiuser detection, 301 noise transfer function, sigma delta converters, 2231, 2231, 2232, 2233-47, 2234 non real time variable bit rate, 551, 1658 non return to zero constrained coding techniques for data storage and, 570-571, 581 magnetic recording systems and, 2250, 2250 microwave and, 2567 non return to zero signals optical signal regeneration and, 1759-1763 partial response signals and, 1933-34, 1933 phase shift keying and, 713 non return to zero inverse (NRZI) compact disc and, 1735 constrained coding techniques for data storage and, 570-571, 579, 580 magnetic recording systems and, 2250, 2250 magnetic storage and, 1327, 1330 optical recording and, 579, 580 partial response signals and, 1933, 1934 non return to zero on off keying, optical transceivers, 1826-27 non zero dispersion shifted fibers, optical fiber systems, 1845, 1848 noncoherent modulation, in underwater acoustic communications, 40-41 noncoherent processing, in acoustic modems for underwater communications, 16 nongeosynchronous earth orbit, broadband, 2656 noniterative algorithms, equalizers, 82 nonlinear components, simulation, 2289-90 nonlinear distortion, RF power amplifiers, compensation, 530-538, 533 nonlinear effects in optical fiber (see also optical fiber), 1683-88, 1711-12, 1845-45, 1876 cross phase modulation and, 1684, 1686-87, 1687, 1712 dispersion compensating fiber in, 1686 dispersion management in, 1686 four photon mixing in, 1687-88, 1687, 1712 refractive index in, 1686 scattering in, 1684-85, 1685, 1712 self-phase modulation and, 1684, 1686, 1686 sidebands and two-tone products in, 1687 solitons and, 1686

stimulated Brillouin scattering in, 1684, 1712

stimulated Raman scattering in, 1684-85, 1685, 1712 stimulated scattering in, 1684-85, 1685, 1712 wave division multiplexing and, 1684 nonlinear least square, wireless systems, 2690 nonlinear optical loop mirrors, optical signal regeneration, 1761, 1762 nonlinear Schrodinger equation, optical communications systems, 1488, 1489, 1491 nonlinearity, in cable modems, 327-328 nonorthogonal spreading, in code division multiple access, 2874-75 nonprimitive BCH coding, 244-252 nonprimitive elements, in BCH coding, binary, 239 nonradiative dielectric, 1243-44, 1244, 1428, 1428 nonradiative dielectric waveguide, 1390 nonreal time VBR, ATM, 206, 267 nonrepudiation, 1151, 1152, 1648 nonuniformly spaced tapped delay line equalizers (see tapped delay line equalizers) Nordic Telecommunications, 2586 Nordstrom-Robinson coding, 890-891 normal response mode, in high level data link control, 546 normalized least mean square algorithm, 7, 8, 9-12 normalized linear density, digital magnetic recording channel, 1324 normalized phase smoothing response., minimum shift keying, 1458 normalized squared Euclidean distance, SCCPM, 2182-84 normalized sum rate, wireless multiuser communications systems, 1608 North American TDMA, 126 north south curve and skywaves, 2061 notarization, 1649 NSFNET, 268 NSTAR, 2112 NTSC standard cable modems and, 326 community antenna TV and, 522 digital versatile disc and, 1738 high definition TV and, 966-979 terrestrial digital TV and, 2546 null beamwidth antenna arrays, 143 numerical aperture, in optical fiber, 435 numerically controlled oscillator, in cable modems, 328 NxN crossbar switching, ATM, 201-203, 202 Nyquist condition, partial response signals, 1929, 1930 Nyquist criterion digital phase modulation and, 709, 710 intersymbol interference and, 1160-61 Nyquist filters, 414, 2460 Nyquist frequencies, in magnetic storage, 2, 1330 Nyquist function, quadrature amplitude modulation, 2049 Nyquist limit, photonic analog to digital conversion, 1960 Nyquist property, cable modems, 328, 333 Nyquist pulses, quadrature amplitude modulation, 2045, 2046 Nyquist rate, 2107 speech coding/synthesis and, 2371 waveform coding and, 2831 Nyquist theorem, 86, 2370 Nystrom's method, in antenna modeling, 173 Oakley, 2813 object-based image and video coding, 1057, 1057 objective function, in adaptive receivers for spread-spectrum system, 99-100, 99 odd-even piggybacking, 233 Odyssey acoustic telemetry, 24 offered load, in traffic engineering, 486 Office of Naval Research, 25 offset orthogonal frequency division multiplexing and, 1875 parabolic and reflector antenna and, 1921, 1921 offset QASK, 2046 offset quadrature amplitude modulation, 2549-55, 2548 offset quadrature phase shift keying, 354, 717, 1459-64, 1460, 1461, 1472

Ohm's law, sound propagation, analogy to, 31, 31 oil exploration acoustic telemetry, 24 Okumura-Hata model, in cell planning in wireless networks, 376 omni sites, in wireless networks, 374 omnicells, 450 omnidirectional antenna, 197-198 on demand multicast routing protocol, 2891, 2892 on off keying code division multiple access, 2731-33 optical synchronous CDMA systems and, 1809, 1813-16, 1815 optical transceivers and, 1826-28 wireless infrared communications and, 2927, 2928 onboard processing (see also communications satellite onboard processing), 880-881, 2113, 2114 onboard switching in satellite communications, 2113 one way functions, cryptography, 606, 609-610, 1152 open loop control, ATM, 551 open loop power control, 1986 open service architecture, IMT2000, 1100-01 open shortest path first, 549-550, 1153, 1533-34, 1534, 2462 ATM and, 204 flow control, traffic management and, 1658 IP networks and, 269 Open Systems Interconnection (see OSI reference model) open waveguide, and active antenna, 51-52 operations and maintenance function, ATM, 200-201, 206 optical add drop multiplexer, 651, 748-749, 751-756, 752, 754, 755, 1727, 1786, 2839-40, 2840, 2864, 2864, 2867-71, 2868, 2871 optical birefringence, optical communications systems, 1492 optical character recognition, Viterbi algorithm, 2818 optical clock extraction, optical signal regeneration, 1762-63, 1763 optical communications, in underwater acoustic communications, 36 optical couplers, 1697-1700, 1697-1700 optical cross connects/switches (see also burst switching networks), 1700-06, 1701, 1702, 1782-97, 1783 acoustooptical tunable switches in, 1785 architecture for, 1800, 1800 arrayed waveguide grating in, 1786-90, 1787, 1788 bandwidth and, 1784, 1797 bit error rate and, 1785 blocking vs. nonblocking, 1783-84, 1783 burst switching network and, 1801-07, 1802 chromatic dispersion and, 1784 circuit switched networks and, 1800 control planes in, 1798 crosstalk and, 1784-85 dense WDM and, 1783, 1797 design constraints on, 1784 electroptical switches and, 1785 external network to network interface in, 1799 fabric technologies for, 1786-1796 fiber delay lines and, 1804-06, 1805 fiber switch capable interfaces for, 1799 generalized MPLS in, 1799 insertion loss and, 1784 internal network to network interface in, 1799 IP networks and, 1798 IP over WDM in, 1798, 1799 label switched path in, 1798-99 label switching in, 1798 lambda switch capable interfaces for, 1799 lightpath topologies in, 1798 liquid crystal switches and, 1790-91 loss and, 1784 Mach-Zehnder interferometer and, 1785 microelectromechanical systems and, 1784, 1785, 1793-96, 1794, 1795 multicast capable OXCs, 2104 multiprotocol label switching and, 1798 optical add drop multiplexers and, 1786 optical electrical optical conversion using, 1782

OSI reference model and, 1798

optical cross connects/switches (see also burst switching networks) (continued) packet switch capable interfaces for, 1799 packet switching and, 1798, 1800-01, 1801 passive optical cross connects and, 1786-90 passive, 1786–90, 1786 photonic analog to digital conversion and, 1961 polarization dependent loss and, 1784 polarization mode dispersion and, 1784 quality of service and, 1798, 1804-06 research in, 1783 reservation protocols and, 1800, 1801 routing and wavelength assignment protocol for, 1800, 2098-2105, 2099 semiconductor optical amplifier and, 1785 SONET and, 1782, 1798 static vs. dynamic WDM networks in, 1798 switching time in, 1785-86 synchronous digital hierarchy and, 1798 thermocapillary switches and, 1792-93, 1792 thermooptic switches as, 1785 time division multiplex and, 1799 transparency of protocols in, 1797 user to network interface in, 1799 waveguide grating router in, 1786 wavelength converters and, 1799 wavelength division multiplexing and, 1797-1808, 2864, 2864, 2867-71, 2868 wavelength routing networks in, 1798, 1799-1800 optical detectors, free space optics, 1857-59, 1858 optical distance profile, sequential decoding of convolutional coding, 2160-61, 2161 optical electrical optical, 1701, 1702, 1782-97 optical fiber systems (see also modeling and analysis of digital optical communication systems; solitons), 434–440, 1706, **1708**, 1714–22 absorption in, 1709, 1710 access networks in, 1840, 1841 acoustic jitter in, 1767 ALOHA protocols and, 1720 amplification in, 1767 amplified spontaneous emission in, 1842, 1843, 1844-45, 1847-48 amplifiers for, 1707, 1709-10, 1710, 1842, 1848 amplitude modulation in, 1826-30 analog to digital conversion in (see also photonic analog to digital converters), 1960-70, 1961 applications for, 434, 1707 asynchronous transfer mode and, 1719, 2615, 2619-20 asynchronrous transmission and, 1808 attenuation in, 1708, 1708, 1709, 1710-11, 1714, 1843, 1844 balanced driving in, 1827 bandwidth in, 436, 1719–20, 1797 bending radius for, 438-439, 438 birefringence in, 1711 bit error rate and, 1841, 1846-47, 1971, 1971, 1972, 1973, 2614 blind equalizers and, 287, 296 Bragg gratings and, 1709 Brillouin scattering and, 1491, 1684, 1712 buffers in, 2614 bus topologies and, 1716, 1716 carrier sense multiple access and, 1808 carrier suppressed return to zero in, 1828-29, 1829 chalcogenide (crystal) glass used in, 434 characterization of, 434–440 chirped return to zero in, 1830 chromatic dispersion and, 1842, 1844, 1845, 1849, 2869 chromatic dispersion in, 436, 1507 circuit switched, 2614–15 circulators in, 1709 cladding in, 434, 435, 1708, 1708, 1714, 1715 coatings for, 434 coherent detectors in, 1848 community antenna TV and, hybrid systems in, 512, 518-522, 518 connectors for, 1707 control planes in, 1798

core in fibers, 434, 435, 1708, 1708, 1714, 1715 core networks in, 1840, 1841 correlation in, digital, 702-709, 703, 705, 708 couplers for, 1707, 1715, 1715 cross phase modulation and, 1684, 1686-87, 1687, 1712, 1844, 1846 crosstalk and, 1759, 1843 cutback method testing of, 436 dark current noise in, 1843 data rates in, 1714 defects and cracks in, intrinsic vs. extrinsic, 437-438 degeneracy factor in, 1846 dense WDM and, 748-757, 749, 1709, 1720-21, 1797 development of, 1484 differential group delay in, 1970-71, 1971 differential phase shift keying in, 1830-31, 1831 differential quadrature phase shift keying in, 1831–1832, **1832** dispersion compensating devices for, 1848 dispersion compensating fiber in, 1686, 1712, 1768, 1846 dispersion in, 436, 1709, 1711, 1686, 1764, 1765 dispersion shifted fiber in, 1711, 1714, 1845, 1848 diversity and, 735 doping in, 1484 dual queue dual bus and, 1715 duobinary and modified duobinary signals and, 1829-30, 1829 dynamic range in, 1718 edge networks in, 1840, 1841 effective length in, 1489 electric fields of, 1488 electro absorption modulated lasers in, 1826 electro absorption modulators in, 1826 erbium doped fiber amplifier in, 1484, 1709, 1721, 1842, 2273, 2839, 2869 Ethernet and, 1507, 1510-11, 1717, 1719 evolution of, 1798, 1798 fatigue testing in, 438-439 fault tolerance and, 1636 fiber distributed data interface and, 1715, 1718-19, 1808 fiber stress history and, 439 fiber to the building and, 1797 fiber to the curb and, 1797 fiber to the home and, 1797, 1808 Fibre Channel and, 1719 forward error correction in, 1848 four photon mixing in, 1687-88, 1687, 1712 four wave mixing in, 1843, 1846 free space optics in (see free space optics) frequency division multiplexing and, 1709 gain in, 1842-43 geometric characterization of, 434-435 Gigabit Ethernet and, 1507, 1509, 1721 Gordon-Haus effect in, 1490, 1767, 1769 group velocity dispersion in, 1764, 1765, 1769 history and development of, 1706-07 index of refraction in, 1488 infrared, 434 inline optical amplifiers in, 1710 insertion loss in, 1843 installation of, 1707-08 interference in, 1484 Internet and, 2461-72 IP over WDM in, 1798, 1799 jitter in, 1767, 1769 Jones vectors and Jones matrix in, 1492 Kerr effect and, 1765 laser intensity noise in, 1843 laser phase noise in, 1843 laser sources for, 1708, 1714, 1842 light emitting diode sources for, 1708, 1714 light sources for, 1708, 1714, 1775-82 lightpath topologies in, 1798 lightwave systems using, 1707 local area networks and, 1714–22, 1808, 1840 losses in, 1710-11, 1767, 2614 Mach-Zehnder interferometers in, 1709, 1826-27 macrobending attenuation in, 439

manufacturing process for, 1708 measurement methodologies for, 1713 mechanical characterization of, 436-440 medium access control and, 1716 metropolitan area networks and, 1714, 1808, 1840 microbending sensitivity in, 439 modal dispersion in, 1507 modal noise in, 1843 modeling and analysis of digital communications systems using, 1484-94 modulation and, 1708, 1825-32, 1848 multimode fiber in, 434, 1507, 1707, 1842 multiple access interference in, 1809 multiprotocol label switching and, 2615 network design considerations for, 1717-18 noise and, 1759, 1824-40, 1843-45, 1847-48 non return to zero on off keying in, 1826-27 non zero dispersion shifted fibers in, 1845, 1848 nonlinear effects in, 1683-88, 1711-12, 1845-46 nonlinear Schrodinger equation in, 1488, 1489, 1491 numerical aperture of, 435 on off keying in, 1826-28 optical birefringence in, 1492 optical bypass in, 1715, 1716 optical cross connects/switches and, 1797-1808 optical multiplexing and demultiplexing and, 1748–59, **1748** optical synchronous CDMA systems and, 1808-24 optical time division multiplexing and, 1828 optical time domain reflectometry testing of, 435-436 optically controlled millimeter wave antenna and, 1431, **1431** OSI reference model and, 1798 packet switching networks and, 1798 passive optical networks and, 1510-12, 1511, 1512, 1717 pathlength in, 1849 phase modulation in, 1830-32 phase shaped binary transmission in, 1829 photocurrent in, 1709 photodetectors for, 993-1006, 1709 photodiodes in, 1842 photonic analog to digital converters in (see also photonic analog to digital converters), 1960 photonic systems using, 1707 pin receiver in, 1832-33, 1832 plastic used in, 434 polarization and, 1491-93, 1711, 1759, 1768 polarization dependent loss in, 1843 polarization maintaining fiber and, 1972 polarization mode dispersion in, 436, 436, 1843, 1845, 1970–75, **1970** power law crack growth method in, 439 power management in, 1847, 1849 power/booster amplifiers for, 1710 preamplifiers for, 1710 principal states of polarization in, 1970-71, 1971 proof test machine for, 437, 437 propagation in, 1765 protocols and, 1718-19 pseudo multilevel or polybinary signals in, 1825 pulse carver in, 1828 pump waves in, 1712 push pull operation in, 1827 Q factor in, 1846-47 quantum efficiency in, 1842 Raman fiber amplifiers for, 1709, 1842 Raman scattering in, 1491, 1684-85, 1685, 1712 receivers for, 1709, 1824-40 refractive index in, 1686, 1715, 1765 regenerators for, 1707 reliability and, 439, 1636, 1636 return to zero on off keying in, 1827-28, 1827 ring topologies and, 1716, 1716 scattering in, 1491, 1709, 1710, 1766, 1843, 1844, 1846 security in, 2614 self-healing ring topologies and, 1716, 1716 self-phase modulation and, 1489, 1489, 1684, 1686, **1686**, 1765, 1844, 1846, 1974

3044 INDEX

optical fiber systems (see also modeling and analysis of digital optical communication systems; solitons) (continued) semiconductor optical amplifiers and, 1826, 1842 shot noise in, 1843 sidebands and two-tone products in, 1687 signal processing in, 1808 signal quality monitoring and, 2269 signal regeneration in, 1759–64 signal to noise ratio and, 1709, 1841, 1846-48 silica glass used in, 434 single mode fiber in, 1507, 1707, 1842, 1845, 1848 solitons and, 1686, 1714, 1764-73, 1848 SONET and, 1798 spectral attenuation in, 435 spectral efficiency in, 1848-49 splices in, 440, 1707 star topologies and, 1716-17, 1717 static vs. dynamic WDM networks in, 1798 stimulated Brillouin scattering in, 1684, 1712, 1844, 1846 stimulated Raman scattering in, 1685-85, 1685, 1712, 1843, 1846 stimulated scattering in, 1684-85, 1685, 1712 Stokes photons in, 1712 storage area networks and, 1714 strength of, 438 switches for, 1782–97, 1783, 1782 synchronous digital hierarchy and, 1798, 2615 synchronous transmission and, 1808 system engineering for (see also optical transport system engineering), 1840-49, 1841 test and measurement of, 2572-79 test procedures for, 434 thermal noise in, 1843 third order dispersion and, 1766-67 time division multiple access and, 1808 topologies using, 1715-17 transmission characterization of, 435-436 transmission control protocol in, 2618-19 transmission using, 1488–91, 1489, 1491 transmitters for (see also optical transceivers), 1707, 1824-40 transport protocols for, 2513–22 very high speed DSL and, 2782–84 virtual LAN and, 1721-22 water's degrading effect on, 437 wave division multiplexing and, 1684 wavelength division multiplexing and, 1709, 1714, 1719–21, 1720, 1759, 1768, 1769, 1808, 1824, 1841, 1841, 2087-2100, 2098, 2614, 2615, 2838-46, 2839 wavelength of, 434, 1714 wavelength routing networks in, 1798, 1799-1800 Weibull fracture probability distribution in, 438 wide area networks and, 1714, 1840 window regions, 1708 optical filters, 1722-33 acoustoopic, 1729-30, 1729 acoustooptic filters in, 1756 arrayed waveguide grating router, 1723, 1724, 1731, 1731 bandwidth and, 1732 Bragg gratings and, 1723, 1723, 1727-28, 1727, 1728 bulk diffraction gratings as, 1723, 1723, 1725-26, 1726 center wavelength in, 1731 characteristics of, 1731-32 control mechanisms for, 1724 counterpropagating gratings (see also Bragg gratings), 1723, 1723 coupled mode theory and, 1728-29, 1729 dielectric thin film stack interference filters, 1723, 1723, 1726-27, 1726, 1749 diffraction gratings, 1723, 1723, 1725-26, 1726 evaluation of, 1732 Fabry–Perot interferometer as, 1723, **1723**, 1724–25, **1724**, **1725**, 1756–57, **1757** figure of merit in, 1731-32 finesse in, 1724 free spectral range in, 1724

insertion loss and, 1732 interference and, 1723 interference filters in, 1756-57 Mach-Zehnder interferometer and, 1723-24, 1723, 1730, 1730, 1757-58, 1757 mode converting (long period) gratings as, 1723, 1723, 1727-28, 1727, 1728 multicavity, 1725 multipass, 1725 optical multiplexing and demultiplexing and, 1756-58 overcoupling in, 1729 polarization and, 1732 scalability and, 1732 selective vs. corrective, 1723 tunable, 1724 Vernier principle in, 1725 waveguide grating filters as, 1723, 1723, 1727-28, 1727, 1728 wavelength division multiplexing and, 1723, 1731-32 optical folding flash AD converter, 1963-64, 1963 optical interfereometers, active antenna, 53 optical Internet, survivable, 2461-72 preconfigured cycle in, 2468-69 protection cycles in, 2469 ring cover/node cover in, 2468 routing in, 2469 self-healing rings and, 2464-68 shared protection in, 2465-66 shared risk link group and, 2463-64 short leap shared protection in, 2466-68 SONET and, 2464-68 spare capacity allocation in, 2468-70 static SLSP in, 2469-70 optical isolators, lasers, 1781 optical line termination, Ethernet, 1511, 1512 optical memories (see also compact disc; digital versatile disk), 1733-41, 1734 blue violet lasers in, 1739 CD-R media in, 1736-37, 1736 CDROM and, 1733-35, 1735, 1736 CD-RW media, 1737 channels in, 1733 compact disc, 1735–36, **1736** digital versatile disc, 1737–38, **1737** disc based, serial type, 1733-35 DVD-ROM media in, 1738 fluorescent discs in, 1739 holographic, 1740, 1740 jukeboxes, 1733 Kerr effect in, 1739 lasers in, 1739 magnetooptic disks in, 1738-40 magnetooptic magnetic field modulation in, 1739 mastering of discs in, 1734 read process in, 1733 receivers for, 1733 recordable DVD-R media, 1738 solid immersion lens technologies in, 1739 standards for, 1736, 1737 storage area networks and, 1733 transmitters for, 1733 vertical cavity surface emitting lasers in, 1739 write once read many, 1737 write process in, 1733 optical modulators, 1741-48 asymmetric coplanar stripline in, 1744 asymmetric stripline in, 1744 automatic bias control in, 1746 bandwidth and, 1744 basic structure and characteristics of, 1742-44 beam propagation method and, 1745 Bessel functions in, 1742 chirp in, 1743-44 coplanar electrode structure in, 1744 DC drift in, 1746-47, 1746 driving point in, 1746 driving voltage in, 1743 driving voltage reduction in, 1745-1746 electrode structures in, 1744-47, 1744 electrooptic (Pockels) effect in, 1742

extinction ratio in, 1743 feedback control in, 1746 insertion loss in, 1743 lasers and, 1741 Mach-Zehnder interferometer in, 1742-44, 1743 modulation in, 1742, 1744 polarization in, 1742 propagation beam method and, 1745 refractive index and, 1745, 1745 reliability in, 1746 thermal drift and, 1747 time division multiplexing and, 1741 wavelength division multiplexing and, 1741 Y branch waveguide in, 1742 optical multiplexing and demultiplexing (see also diffraction gratings), 1748-59, 1748 acoustooptic filters in, 1756 acoustooptical gratings in, 1755-56, 1756 arrayed waveguide grating in, 1752-54, 1753 Bragg gratings and, 1749 bus architecture in, 1758 coupling loss and, 1751-52, 1753 crosstalk and, 1752 diffraction gratings in, 1749-56, 1758 diffraction in, 1749 erbium doped fiber amplifiers and, 1748 example and assessment of, 1758 Fabry-Perot interferometers in, 1749 free space gratings in, 1754-55 functionality in, 1749 interference in, 1749 Mach-Zehnder inteferometers in, 1749 Michelson interferometers in, 1749 microelectromechanical switches in, 1758 optical fiber systems and, 1748-59, 1748 optical filters in, 1756-58 optical time division multiplexing and, 1748 photonic analog to digital conversion and, 1964-65 Sagnac interferometers in, 1749 thin film stack interference filters, 1749 wavelength division multiplexing and, 1748 wavelength routing in, 1749 optical network unit, 1511, 1512, 2780-81 optical networks, medium access control protocols, 1551-62 optical orthogonal coding, 1809, 2730-31 optical packet switching networks, 1798 optical parametric oscillator, 1853 optical receivers (see also optical transceivers), 1824-40 optical recording, 1319 combicoding in, 581 DC control in, 579-581, 580 frequency domain constraints in, 579 non return to zero in, 581 non return to zero inverse in, 579, 580 power spectral density function in, 579 run length limited in, 579-581 running digital sum in, 579 spectral null constraints in, 579, 580 substitution coding in, 581 time domain constraints in, 579 optical signal regeneration, 1759-64 all optical, 1759 before and after, 1760 counterdirectional scheme for, 1761 crosstalk and, 1759 decision characteristics in, 1760 delayed interference devices in, 1763, 1763 differential delay in, 1761, 1761 distributed feedback lasers in, 1762 electroabsorption modulators in, 1761-62 extinction ratio improvement in, 1760 figures of merit in, 1759-60 Mach-Zehnder interferometer in, 1760-61, 1760 Michelson interferometer in, 1760-61, 1760 mode locked lasers in, 1762 multichannel, 1763-64, 1763, 1764 multimode interference coupler in, 1761 noise in, 1759 noise reduction in, 1760 non return to zero signals and, 1759-1763

signal paths in, 1842

optical signal regeneration (continued) nonlinear optical loop mirrors in, 1761, 1762 optical clock extraction in, 1762-63, 1763 optical waveguide and, 1760 optoelectric, 1759 phased arrays in, 1763 photonic integrated circuits in, 1759 polarization control in, 1759 Q-switched lasers and, 1762, **1762** reamplification in, 1759 reshaping in, 1759 return to zero signals and, 1759-1763 semiconductor optical amplifier based, 1760-63, 1760 3R regeneration in, 1762-63, 1763 timing jitter in, 1759 tunable bandpass filters in, 1764 2R regeneration in, 1760-62, 1760 wavelength conversion in, 1760 wavelength dependent couplers in, 1761 wavelength division multiplexing and, 1759 optical sources (see also lasers; light emitting diodes) history and development of, 1775 lasers as, 1776-81 light emitting diodes as, 1775-76 solitons and, 1770 wavelength division multiplexing and, 1775 optical storage (see also optical memories), secure ultrafast data communication/processing in (see also holographic memory), 2132-40, 2133 optical switching techniques in WDM optical networks, 1797-1808 optical synchronous CDMA systems, 1808-24 coding acquisition in, 1813 coding tracking in, 1813 encoding in, 1809 frequency encodingd CDMA in, 1816-17, 1817, 1818 frequency encodingd systems in, 1816-17, 1817, 1818 Gaulois fields in, 1810 intensity modulation/direct detection in, 1809, 1814 interference cancellation in, 1817-23, 1819-23 liquid crystal modulators in, 1817 modified prime coding in, 1812 modulation in, 1809, 1813 multiple access interference in, 1809, 1810 on off keying and, 1809, 1813-16, **1815** optical orthogonal coding in, 1809 prime coding in, 1810 pseudoorthogonal coding for, 1809-10 pulse position modulation in, 1809, 1813, 1815-16, 1816 quasiprime coding in, 1811 receivers for, 1815, **1815** sequences for, 1809 tapped delay lines in, 1815, 1815, 1816 time encodingd systems in, 1813-16 transmitter for, 1815-16, 1816 two/2n prime coding in, 1811-12 optical time division multiplexing, 1748, 1828 optical time domain reflectometry, 435 optical transceivers, 1824-40 amplitude modulation in, 1825, 1826-30 avalanche photodetection in, 1834 bandwidth optimization in, 1836-37 beat noise in, 1836 carrier suppressed return to zero in, 1825, 1828 coherent detection in, 1834-35, 1834 differential phase shift keying in, 1825, 1830-31, 1831 differential quadrature phase shift keying in, 1831–1832, **1832** electro absorption modulated lasers in, 1826 electro absorption modulators in, 1826 electronics noise in, 1833 equivalent noise current density in, 1833 erbium doped fiber amplifiers and, 1835 frequency modulation in, 1825 heterodyne receivers in, 1835 homodyne receivers in, 1835 Mach-Zehnder interferometers in, 1826-27 noise equivalent power and, 1833 non return to zero on off keying in, 1826-27

optical time division multiplexing and, 1828 optically preamplified detection in, 1834, 1835-36, 1835 phase amplitude shift signaling in, 1829 phase modulation in, 1830-32 photonic integrated receivers in, 1838 pin receiver in, 1832-33, 1832 polarization in, 1825 polarization mode dispersion and, 1825 Q factor in, 1825, 1832 quadrature amplitude modulation in, 1825 quantum limit in, 1833, 1837 return to zero DPSK in, 1825 return to zero on off keying in, 1827-28, 1827 semiconductor optical amplifiers and, 1826 shot noise limit in, 1835 sidebands in, 1826 signal to noise ratio in, 1825, 1837 transimpedance in, 1833 optical transmitters (see also optical transceivers), 1824-40 optical transmitters, receivers and noise (see optical transceivers) optical transport system engineering access optical networks and, 1840, 1841 amplified spontaneous emission in, 1842, 1843, 1844-45, 1847-48 amplifiers in, 1848 attenuation in, 1843, 1844 bit error rate in, 1841, 1846-47 chromatic dispersion and, 1842, 1844, 1845, 1849 coherent detectors in, 1848 core optical networks in, 1840, 1841 cross phase modulation in, 1844, 1846 crosstalk in, 1843 dark current noise in, 1843 degeneracy factor in, 1846 dispersion compensating devices for, 1848 dispersion compensating fiber in, 1846 dispersion shifted fiber in, 1845, 1848 edge optical network in, 1840, 1841 enabling technologies and tradeoffs in, 1848 erbium doped fiber amplifiers in, 1842 extinction ratio in, 1842 fiber type selection in, 1848 forward error correction in, 1848 four wave mixing in, 1843, 1846 gain in, 1842-43 impairment parameters in, 1843-44 insertion loss in, 1843 laser chirp in, 1844 laser intensity noise in, 1843 laser phase noise in, 1843 lasers in, 1842 limitations and penalties, assessment of, 1844-46 local area networks and, 1840 metropolitan area networks and, 1840 modal noise in, 1843 modulation and, 1848 multimode fiber in, 1842 noise accumulation in, 1847-48 noise and, 1844–45, 1847–48 noise parameters in, 1843 non zero dispersion shifted fibers in, 1845, 1848 nonlinear effects in, 1845-46 optical amplifiers for, 1842 output power in, 1842 pathlength in, 1849 photodiodes in, 1842 polarization dependent loss in, 1843 polarization mode dispersion in, 1843, 1845 power penalty handling in, 1847, 1849 Q factor in, 1846–47 quantum efficiency in, 1842 Raman amplifiers in, 1842 responsivity of photodiodes in, 1843 self-phase modulation in, 1844, 1846 semiconductor optical amplifiers in, 1842 shot noise in, 1843 signal parameters for, 1842-43

on off keying in, 1826-28

signal to noise ratio in, 1841, 1846-48 single mode fiber in, 1842, 1845, 1848 solitons and, 1848 spectral efficiency in, 1848-49 stimulated Brillouin scattering in, 1844, 1846 stimulated Raman scattering in, 1843, 1846 thermal noise in, 1843 transmission parameters in, 1840-44 wavelength division multiplexing and, 1841, 1841 wide area networks and, 1840 optical waveguide, optical signal regeneration, 1760 optical wireless laser communications (see free space optics) optically controlled millimeter wave antenna, 1431, 1431 optically preamplified detection, optical transceivers, 1834, 1835-36, 1835 optimal detectors, in adaptive receivers for spreadspectrum system, 98-99 optimal digital filters, 699 optimal path determination, 549 optimal receivers, for adaptive receivers for spreadspectrum system, 102-103 optimal selection diversity, 731 optimal zero memory nonlinear devices, 2414 optimization, 372 antenna arrays and, 160-164 cell planning in wireless networks and, 382-383 quantization and, 2130 underw/in underwater acoustic communications, 45 optimized link state routing, 2889 optocouplers, in lasers, 1781 optoelectric signal regeneration, 1759 optoelectronic regenerators, 2863 Orange Book, 1737 Orbcomm satellite communications, 1251 orbit of satellite, 877, 1248-49, 1248, 2113 organizationally unique identifiers, 1282 orthogonal coding, Golay complementary sequences, 896-898 orthogonal frequency division multiple access, 321, 322, 1878 orthogonal frequency division multiplexing, 1873, 1867–79, **1867** adaptive loading in, 1878, 1878 additive white Gaussian noise in, 1874 adjacent channel interference in, 1876 applications for, 531-532, 1878 autocorrelation and, 1945 baseband and passband representations in, 1872 baseband PAR in, 1945 basic technique for, 1868-71 binary phase shift keying in, 1945, 1948 bit interleaved coded modulation and, 278 broadband wireless access and, 320, 321, 322 channel estimation and correction in, 1877-78 channel time variations in, 1875 clipping and, 1946-47 coding divison multiple access and, 1878 coding for, 1873, 1876 coding rate in, 1947, 1947 coherent detection in, 1877-78 complementary cumulative distribution functioin in, 1945-46, 1946 copper media and, 1867 cumulative density function in, 1946 cyclic extension in, 1872, **1872** data frames in, 1945 detection techniques in, 1877 differential detection in, 1877 digital audio/video broadcasting and, 678-679, 1867, 1878 digital to analog conversion in, 1871 discrete Fourier transform and, 1944 discrete multitone and, 736, 1878 discrete multitone transmission in, 1944 dispersion in, channel time, 1874 envelope of, 1869-70 envelope power function in, 1948 equalizers and, 93

orthogonal frequency division multiplexing (continued) expectation maximization algorithm and, 773 factors of merit in, 1951 fast Fourier transform in, 1871 filters in, 1871-72 Fourier transforms for, 1869-70 frequency division multiple access and, 828 frequency domain equalization in, 1877 frequency offset in, 1875 frequency response in, 1874 frequency spectrum for, 1868-71 gain in, 1875 Golay complementary sequences and, 893 guard interval in, 1872 history and development of, 1867 impairments to channel and system in, 1874-76 interference and, 1874, 1876 intersymbol interference and, 1867 inverse discrete Fourier transform in, 1871-72 Kineplex and Kathryn systems for, 1867 mobile radio communications and, 1867 multicarrier CDMA and, 1521, 1523, 1524, 1525-28, 1526 multicarrier transmission in, 1867-68, 1868 multiple input/multiple output systems and, 1456, 1878 noise and, 1874–76 passband PAR in, 1945 peak to average power ratio in, 1876 peak to average power ratio in, 1944-53 peak to mean envelope power ratio in, 1945 phase noise in, 1875 phase shift keying and, 1945 pilot patterns for, 1877, 1877 power spectral density of, 1869-70, 1870, 1873, 1873 powerline communications and, 1995, 2001, 2003 predistortion/compensation in RF power amplifiers and, 530-532, 535 primary spectrum in, 1871 quadrature amplitude modulation and, 1868 quadrature phase shift keying and, 1869, 1945, 1947 RAKE receivers and, 1878 receivers for, 1867-71 sample design using, 1873-74, 1873 scrambling in, 1876 signal constellations in, 1945 signal distortion in, 1876 signal to interference plus noise ratio in, 1874 signal to noise ratio in, 1873 software radio and, 2314 space-time coding and, 2328-29, 2329 subcarriers/subchannels, 1867 synchronization and, 2481-85, 2481, 2482 terrestrial digital TV and, 2549 timing errors in, 1875-76 transmitter nonlinearities in, 1876 transmitters for, 1867-71, 1869 in underwater acoustic communications, 41 unequal error protection coding and, 2766-67 windowing in, 1872-73, 1873 wireless communications, wireless LAN and, 1288-89, 1867, 2916, 2941-45, **2944**, **2945**, **2946** wireless transceivers, multi-antenna and, 1582 orthogonal method, in antenna arrays, Fourier transforms, 157–158, 158 orthogonal multistage filters, packet rate adaptive mobile receivers, 1892, 1892, 1893 orthogonal perturbation, in antenna arrays, 159, 159, 160 orthogonal signaling, minimum shift keying, 1457 orthogonal spreading in code division multiple access, 2874-75 orthogonal time division, cdma2000, 361, 367 orthogonal variable spreading factors, 387 orthogonality of signals, in code division multiple access, 458 orthogonality restoring detector, multicarrier CDMA, 1527 orthosynthesis antenna arrays, 158-159, 159 oscillators, 1478

active antenna and, 51, **51**, 52, 58–60, **58**, **59**, **60**, 63, 65, 66, **66**

cable modems and, 328 frequency synthesizers and, 837-843, 838-843 surface acoustic wave filters and, 2454-55, 2454 OSI reference model, 15-16, 539-540, 539, 540 automatic repeat request and, 225-226 cdma2000 and, 359, 365 Ethernet and, 1502-05, 1502 optical cross connects/switches and, 1798 optical fiber and, 1798 packet switched networks and, 1910-12, 1911 satellite communications and, 2118 satellite onboard processing and, 482-483 security and, 1647-50 in underwater acoustic communications, 45 out of band signaling, transport protocols for optical networks, 2618 outage probabilities, cochannel interference and, 451-452 diversity and, 729 intelligent transportation systems and, 503-504, 504 wireless and, 2922 outer coding, 2164 outer loop power control, 1986 output buffered switches, ATM, 201 output contention, ATM, 201 output feedback cryptography, 607 overcoupling, in optical filters, 1729 overhead levels, SONET, 2488-93 overmodulation, amplitude modulation, 134 overprovisioning, multimedia networks, 1563, 1567 oversampling blind equalizers and, 287, 288-289 community antenna TV and, 522 photonic analog to digital conversion and, 1960, 1961, 1965-68, 1965 overspread, cellular communications channels, 395 oxygen and absorption, millimeter wave propagation, 1437, 1437 P median problem in quantization, 2128 P1, P2, P3, Px polyphase sequences, 1977 Pacheco, Ryan A., 2333 packet binary convolutional coding, 2946 packet classifiers, in flow control, traffic management, 1656 packet communications, 15-16 packet data channel control function, cdma2000, 359, 363, 364 packet data protocol, 126, 867 packet demand assignment multiple access, 1347 packet dropping, 1565-66, 1659 packet error rate, 881, 1886, 1887, 1900-01, 1901 packet ID, cable modems, 324 packet loss in broadband, 2655 packet market/tagging, flow control, traffic management, 1659 packet radio networks, 1342-49, 2212 packet rate adaptive receivers adaptive receivers and, 1886 additive white Gaussian noise and, 1886-88, 1901 ALOHA protocols and, 1902-03 angle of arrival estimation in, 1899-1900 autocorrelation in, 1889-90 auxiliary vector filters for, 1890-96, 1891, 1893 basic signal model in, 1887-88 bit error rate and, 1886, 1887, 1892, 1898, 1902 capacity in, 1901-02, 1901 conditional statistical optimization in, 1892 cross validated minimum output variance rule in, 1895, 1898 data packet structure in, 1899, 1899 direct sequence CDMA and, 1886-87, 1886, 1894 error detection and correction in, 1886 fading and, 1886 filtering in, 1888-1900, 1893 forward error correction in, 1887, 1902 frequency division multiple access and, 1886 gain in, system processing gain, 1888 generalized sidelobe canceler in, 1889-90, 1889, 1892-93 interference and, 1886

intersymbol interference and, 1887, 1899 known channel adaptive filter estimation in, 1892-98, 1893 least mean square in, 1883, 1886, 1887 maximum J divergence in, 1895-96 midamble in, 1899 minimum mean square error and, 1886-1903 minimum variance distortionless response in, 1886–1903, 1887 multipath fading and, 1886 multiple access interference and, 1886, 1887 noise in, 1886 packet error rate in, 1886, 1887, 1900-01, 1901 performance of, 1997-98, 1898 quality of service and, 1887, 1901 RAKE processing/RAKE receivers in, 1886, 1887, 1898, 1900, 1901 receiver for, 1887-88, 1887 recursive least square in, 1883, 1886, 1887 sample matrix inversion in, 1886, 1887, 1892-1903 signal to interference plus noise ratio and, 1886, 1895, 1898, 1902 signal to noise ratio, 1898 space-time sequence in, 1886, 1888 spread spectrum and, 1886 subspace channel estimation in, 1899-1900 throughput in, 1902–03, 1902 time division multiple access and, 1886 unknown channel estimation and, 1898-1900 packet reservation multiple access, admission control, 123 packet salvage, ad hoc wireless networks, 2888 packet scale rate guarantee, differentiated services, 674 packet schedulers, flow control, traffic management, 1656 packet switch capable interfaces, optical cross connects/switches, 1799 packet switched networks admission control and, 122 application layer in, 1911 applications for, 1906 asynchronous transfer mode and, 1909 ATM, 200–207, 200 autonomous systems in, 1913 bandwidth and, 1906–07, 1908 best effort networks and, 1910 Bluetooth and, 312–313, 312 bursty transmissions and, 1906 chann/in channel modeling, estimation, tracking, 398-408 circuit switched networks vs., 1906, 1906 classless interdomain routing in, 1912 congestion control in, 1907, 1910 connection oriented vs. connectionless networks in, 1909-10, 1909 data link layer in, 1910-11 domain name servers and, 1913 effective bandwidth in, 1908 Ethernet and, 1910 exterior gateway protocol in, 1913 fault tolerance and, 1632, 1639-40 flow control and, 1625, 1911-12 forwarding in, 1909-10 general packet radio service and, 869 history and development of, 1913-14 interior gateway protocol in, 1913 Internet and, 1912 Internet protocol (IP) and, 1911 IP addressing in, 1912, 1912 IP networks and, 267-271, 1910, 1912 IP telephony and, 1178-79 medium access control and, 1551-55 messages in, 1907 multiplexing in, 1907-09 multiprotocol label switching and, 1909 network layer in, 1911 optical cross connects/switches and, 1782, 1800-01, 1801 OSI reference model and, 1910-12, 1911 physical layer in, 1910 presentation layer in, 1911

packet switched networks (continued) protocols and layering in, 1910-12 reliability and, 1632, 1639-40 resource allocation in, 1908 routers in, 1907, 1909-10, 1913 satellite communications and, 1255 satellite onboard processing and, 482-483 scheduling in, 1908-09 self-healing property of, 1910 session layer in, 1911 SONET and, 1910 statistical multiplexing in, 1907-09, 1908 store and forward networks in, 1907 TCP/IP and, 1912 timescales in, 1908 traffic engineering and, 500 transmission control protocol and, 1911, 1912, 2603 transport layer in, 1911 user datagram protocol in, 1911 wireless, 371, 2981-90, 2983 packet time, in traffic engineering, 500 packets ATM and, 200, 264 Bluetooth and, 312-313, 312 cable modems and, 324 IP networks and, 267 padding, traffic, 1646, 1649 page mode, Bluetooth, 311-312 paging and registration in mobile networks, 311-312, **311**, 366, 1914–28 addressing in, 1914 call to mobility ratio in, 1917 entropy of location distribution in, 1917-18 forwarding in, 1914-15 location determination in, 1915-17, 1915, 1916 mobile agents and, 1918 mobility indexes in, 1917, 1919 mobility management and future of, 1918 point of network attachment and, 1915 primitives and, 1915 quality of service, 1916 random walk in, 1918 theory abstractions of, 1917-18 unit centric approach to, 1916 universal phone numbers in, 1917-18 Web crawlers and, 1915 paging channels, in IS95 cellular telephone standard, 349, 352 pagoda broadcasting, 236 pairing, Bluetooth, for security, 316 pairwise nearest neighbor in quantization, 2128, 2129-30 PAL standard cable modems and, 326 digital versatile disc and, 1738 high definition TV and, 966-979 terrestrial digital TV and, 2546 palette generation problem in quantization, 2128 palindromic coding, 1540 PAR reduction tones, peak to average power ratio, 1952 parabolic and reflector antenna, 1920-28, 1920 aperture and, 2080-81 aperture efficiency in, 1923-24 applications for, 1927-28, 2080 beam deviation factors in, 1926, 1927 beam direction in, 2080 beam in, 1925-26 beam scanning and, 2084-86 blockage of apertures in, 2084 blocked apertures in, 2081 blocking efficiency in, 1923 Cassegrain, 1920–21, **1921**, 2083–86, **2083**, **2084** collimation in, 2082 copolar and cross polar patterns in, 1923 cross polarization efficiency in, 1923 design of, 1921-22 diffraction and, 1924 efficiency in, 1923-24 equivalent parabola in, 1922 far field (Fraunhofer region) radiation concepts in, 2080-81, 2080

feed loss efficiency in, 1924 feed or illuminator in, 1920 feeds for, 1924, 2082, 2083, 2084 field of view in, 1924 focal axis in, 1920 focal length in, 1920, 2084 frequency range for, 1920 gain in, 1923-24, 2080-81 gain to temperature ratio in, 1927 geometric optics analysis in, 1920, 2081-82 Gregorian, 1920–21, 1921, 2083–84, 2083 half power beamwidth in, 1922-23, 1925-26 illumination efficiency in, 1924 lens antenna and, 2082 losses in, 1924 magnification in, 1922 microwave relay links using, 1927-28 mounting, 1924-25 multiple type, 2083-84 noise temperature in, 1922, 1926-27 offset system in, 1921, 1921 parameters and characteristic of, 1920-27, 1922 parameters for, 2082 phase efficiency in, 1923 phased arrays and, 2082 pointing and pointing error in, 1925–26, **1926** power flow, power density in, 2082 primary focus in, 1920 prime focus type, 2082-83, 2082 radation patterns in, 1922-23, 1922, 2080-81, 2081 radio astronomy using, 1927 reflection and, 2082, 2086 remote sensing using, 1928 resolution in, 1923 return loss in, 1924 root mean square value of surface in, 1924 satellite communications using, 1928 secondary focus, secondary mirror in, 1920 shaped systems in, 1921 shaping of dual reflectors in, 2084 sidelobes in, 2081 single type, 2082-83, 2082 Snell's laws of reflection and, 2082, 2086 spillover efficiency in, 1923–24 structural and mechanical aspects of, 1924–25, **1925** subreflector in, 1920, 2083-84, 2083 surface efficiency in, 1924 surface loss efficiency in, 1924 vertex in, 1920, 2083 very long baseline interferometry and, 1927 parallel concatenated coding, 2164, 2179, 2180 parallel concatenated convolutional coding, 2710 parallel entry systolic priority queue, 2149 parallel interference cancellation, in wireless multiuser communications systems, 1617-18 parallel plate capacitor, in active antenna, 49-50, 49 parallel transmission, 1494–95 parasitic elements, in microstrip patch antenna, 1367–68, **1368** Pareto distribution, Pareto exponent, 2157-58, 2157 Pareto optimality criteria, in wireless networks, 372 parity bits, 225, 545, 1540 parity check coding constrained coding techniques for data storage and, 581 low density, 1308-18 multidimensional coding and, 1543-44 threshold decoding and, 2579-85 parity check equation, 2007, 2580-81 parity coding magnetic recording systems and, 2257-58 magnetic storage and, 1331–33 multiple input/multiple output systems and, 1456 park mode, Bluetooth, 314 Park–Park–Song–Suehiro polyphase sequences, 1979 Parlay, 727, 1100 partial band interference, 1130-41 partial distortion search, vector quantization, 2126 partial packet discard, flow control, traffic management, 1660 - 61

partial response magnetic recording systems, 2248

partial response continuous phase chirp modulation, 445, 446 partial response maximum likelihood constrained coding techniques for data storage and, 582 hard disk drives and, 1321 magnetic recording systems and, 2248, 2253-65, 2260 magnetic storage and, 1328, 1328, 1330-31 partial response signals and, 1930-31, 1935 Viterbi algorithm and, 2818 partial response signals, 1928-35 alternate mark inversion in, 1934 bandwidth and, 1929, 1932-33 binary bipolar with n zero substitution in, 1934 channel coding in, 1933 composite response in, duobinary pulse, 1930-32, 1931 detection of, 1931-32 error correction coding in, 1933 faster than Nyquist scheme in, 1932 filtering in, 1928 frequency division multiplexing in, 1929 frequency range of, 1929 full response signals vs., 1928., 1929, 1929 high density bipolar 3 and, 1934 history and development of, 1935 intersymbol interference and, 1928-35 line coding in, 1933-34, 1933 maximum likelihood sequence estimation and, 1932, 1933 modulation in, 1928, 1933-34 noise in, 1928 non return to zero and, 1933-34, 1933 non return to zero inverted in, 1933, 1934 Nyquist condition and, 1929, 1930 partial response maximum likelihood and, 1930-31, 1935 precoding in, 1931-32 pulse amplitude modulation and, 1928, 1933 quadrature amplitude modulation in, 1928 raised cosine spectrum in, 1929-30, 1930 reduced state sequence estimation in, 1933 research in, 1935 return to zero in, 1933-34, 1933 run length limited in, 1934 single sideband modulation and, 1930 symbol by symbol detectors in, 1932 Viterbi algorithm and, 1932, 1933 particle displacement and particle velocity, 31-32 partition-based approach to training in quantization, 2129 partition cells, in transform coding, 2599 Pasquier's construction, Golay coding, 888-889 passband filters, orthogonal frequency division multiplexing, 1872 passband PAR, orthogonal frequency division multiplexing, 1945 passband signal, in channel modeling, estimation, tracking, 399 passband transmission, adaptive equalizers, 80, 80 passive attacks, 1645-46 passive intermodulation effects, 191 passive optical networks, 1717 dynamic bandwidth allocation in, 1511 Ethernet and, 1510-12, 1511, 1512 optical line termination in, 1511, 1512 optical network unit in, 1511 point to multipoint operation in, 1511-12 password security, 1165 PASTd algorithm, 301 patch antenna, 52–53, **52**, 53, 61–62, **61**, 169, 180, 193 patching, 233–234, **234** path accuracy, in radiowave propagation, 2064 path gain factor (propagation factor), 209, 1936 path loss, 781-782 cellular telephony (see also path loss prediction in cellular telephony), 1936-44 indoor propagation models and, 2015 power control and, 1983 satellite communications and, 1223, 1225-26, 1225

path loss prediction in cellular communications, 1936-44 buildings and residential environments, propagation over, 1939-41, 1940 delay and, 1937 diffraction and, 1940 fading in, 1937 geometric optics laws and, 1936, 1942 Keller cone for diffraction in, 1942 line of sight transmission and, 1939, 1941 macrocells and, 1940-41 Maxwell's equations and, 1936 measurement-based models for, 1941 path gain vs., 1936 3D rays for site specific prediction in, 1941-43, 1942 Q factor and, 1939-41, 1940 radiowave propagation and, 216-217 range dependence and microcells in, 1941 ray concepts in, 1936-38, 1936 shadowing in, 1937 shooting-bouncing ray approach to, 1942 sliding averages in, 1937–38 time delay profile of pulsed signal and, 1937 two ray model for flat earth in, 1938-39, 1938, 1939 uniform theory of diffraction and, 1936 vertical plane launch approximation in, 1942-43 path mapping, routing and wavelength assignment in WDM, 2101 path metrics of Viterbi algorithm, 600 PATH project, intelligent transportation systems, 503 pathlength, 2064, 2102 pattern classification/association in neural networks, 1677 pattern diversity antenna, 190 pattern function of elements, antenna arrays, 142 pattern synthesis, in antenna arrays, 153-157 payload mapping, SONET, 2493 payload pointer, SONET, 2489-92, 2492 PCS, 370 peak cell rate, ATM, 266, 551, 552, 1656, 1658 peak detection, magnetic storage, 1327, 1332 peak signal to noise ratio, ADSL, 1576-77, 1576, 1577 peak to average power ratio, 1876 peak to average power ratio asymptotic results in, 1946-47 autocorrelation and, 1945 baseband, 1945 binary phase shift keying in, 1945, 1948 clipping and, 1946-47 coding rate in, 1947, 1947 coding techniques for, 1950-52 complementary cumulative distribution functioin in, 1945–46, **1946** computational methods for, 1947-48 continuous time, 1948 cumulative density function in, 1946 data frames in, 1945 discrete time, 1947 envelope power function in, 1948 Gaussian approach to, 1946 Golay coding and, 1950 Golay-Davis-Jedwab coding in, 1951 infinity norm method in, 1947-48 interleaving approach vs., 1949-50 inverse discrete Fourier transform and, 1947, 1950-51, 1951 merit factor and, 1951 multiple signal generation vs., 1948-50, 1948 PAR reduction tones in, 1952 partial transmit sequences vs., 1949, 1949 passband, 1945 peak to mean envelope power ratio in, 1945 phase shift keying and, 1945 predistortion/compensation in RF power amplifiers and, 530, 532 quadrature phase shift keying in, 1945, 1947 reduction methods for, 1948-52 Reed-Muller coding and, 1950 Rudin-Shapiro recursion in, 1951-52 selected mapping approach vs., 1949 sequence family construction for low PAR/high distance, 1950

signal constellations in, 1945 single sequences with low PAR in, 1950-51 space-time coding and, 2330 system model for, 1950, 1950 peak to mean envelope power ratio, 893, 1945 peak to peak cell delay variation, ATM, 266 peakedness, in traffic engineering, 490-491 peer entities, 540 peer entity authentication, 1647 peer networks, 268, 2208 penetration, in indoor propagation models, 2013, 2018 penetrometer using underwater acoustic modem, 19-20, 20 penultimate hop popping, multiprotocol label switching, 1595 per domain behavior, DiffServ, 675 per hop behavior differentiated services and, 1657, 1658 DiffServ and, 270 multimedia networks and, 1568 per survivor processing, in channel modeling, estimation, tracking, 416, 418 perceptrons, in neural networks, 1676, 1678 perceptual analysis measurement system, 1179 perceptual error weighting, in speech coding/synthesis, 2345 perceptual speech quality measure, 1179, 1305, 2354 perfect electric conductor, 183-184 perfect root of unity sequences, 2330 performance management, 207, 2293-94 perigee in orbit, 1248 periodic broadcasting, 235-236, 236 periodic buffer reuse with thresholding, 234 periodic complementary sequences, 898-899 periodic dielectric millimeter wave antenna, 1428, 1428 periodic impulse noise, powerline communications, 2002 periodically stationary process in the wide sense, 1989-90 permanent virtual circuit, ATM, 204, 265-266 permeability, in antenna modeling, 171 permittivity, in antenna modeling, 171 permutation, in turbo coding, 2711 permutation-based pyramid broadcasting, 236 permutation coding, 1953-60 additive white Gaussian noise, 1954 applications for, 1959 decision regions in, 1955-56 decoding, 1955 definitions in, for calculation, 1954-55 error detection and correction in, 1955-56 evaluation of, 1956-59 maximum likelihood in, 1953, 1954 modulation and, 1953 optimized, 1956, 1956, 1957, 1958, 1959 pulse coding modulation and, 1954 pulse position modulation in, 1954 theory of, 1954-56 variants of, 1953-54 perpendicular recording channels, low density parity check coding, 658-668 Perron-Frobenius theory, in constrained coding techniques for data storage, 574 person in the middle (see man in the middle) personal access communications system, 2952-55 personal area network, 2677, 2682-84 Bluetooth and, 2682, 2683-84 BodyLAN and, 2682 HiperLAN and, 2682, 2683, 2684 Home RF and, 2683, 2684 IrDA and, 2682 standards for, 2683 wireless, 502, 508, 1602 personal communication systems, 7, 350, 1345, 1479, 2306 personal digital assistants, 307, 2190, 2191, 2194, 2314, 2899 personal digital cellular, 194, 195, 2673 personal handyphone system, 2953 personal video recorder, 1319 Peterson's direct solution method, 248–250, 255–257, 260.617

phase synchronization, 2477-78 phase ambiguity problem in channel modeling, estimation, tracking, 416 phase amplitude shift signaling, in optical transceivers, 1829 phase coherent (synchronous) demodulator, 134 phase coherent detection, in underwater acoustic communications, 41-43, 42 phase conjugates in active antenna, 65 phase detector, in frequency synthesizers, 844-845 phase locked loop constrained coding techniques for data storage and, 571 frequency synthesizers and, 830, 832-833, 833, 845-854, **848** pulse amplitude modulation and, 2027, 2027 quadrature amplitude modulation and, 2053-55, 2055 shallow water acoustic networks and, 2207 phase modulation (PM), 807-825, 1830-32, 2179 phase noise, in orthogonal frequency division multiplexing, 1875 phase shaped binary transmission, 1829 phase shift keying (see also digital phase modulation), 709–719, 1335, 2179 additive white Gaussian noise, 712 bit error rate in, 713-15 chirp modulation and, 441, 444 continuous phase frequency shift keying and, 593 differential PSK, 715–717, **716** differential QPSK, 717–718, **718** Gray coding and, 715 high frequency communications and, 947 M ary phase shift keying in, 710-711 modems and, 1497 modulation/demodulation in, 712-713, 712, 713, 714 non return to zero in, 713 offset QPSK, 717 orthogonal frequency division multiplexing and, 1945 peak to average power ratio and, 1945 power spectra of digitally modulated signals and, 1989-91, 1991 power spectral density in, 711-712 predistortion/compensation in RF power amplifiers and, 530 satellite communications and, 1225 serially concatenated coding and, 2173, 2173 shallow water acoustic networks and, 2207 signal constellation in, 711, 711 signal quality monitoring and, 2273 signal to noise ratio, 714 signal waveform in, 710-711 space-time coding and, 2326, 2326 spread spectrum and, 2397 synchronization and, 2473-85 ternary sequences and, 2536-47 trellis coded modulation and, 2622-35 in underwater acoustic communications, 41, 43, 45 phase shifter antenna arrays, 166 phase trellis, 597, 1458-59, 1459, 1467-68, 1468 phased microstrip/microstrip patch antenna and array, 1384-85, **1384**, **1385** phased arrays, 1513-21, 1514, 1763, 2082 phone appliances, home area network, 2685-88, 2687 phonemes, in speech coding/synthesis, 2360-63, 2370, 2371 photo refractive information storage materials, 1740 photoconductors, 1000, 1000, 1431, 1431 photocurrent, in optical fiber, 1709 photodetectors, 993-1006 absorption and, 995, 995 avalanche, 1002 bandwidth and, 1000 capacitance limit in, 999 charge trapping in, 1000 classification of, 1000-01 diffusion limit in, 999-1000 Fabry-Perot interferometer and, 1003 materials for, 995-996 metal semiconductor metal photodetectors in, 1001-02 noise in, 996-997
photodetectors (continued) optical fiber and, 1709 performance and limitations of, 997-1000 photoconductors and, 1000, 1000 photovoltaics and, 1000, 1000 PIN detectors and, 1001 quantum efficiency in, 995, 995 RCE, 1002–1003, 1003 responsivity in, 996 Schottky, 1002, 1002 signal quality monitoring and, 2271 signal to noise ratio, 997 transit time limit in, 998-999, 998 waveguide, 1003-04 wavelength and, 994, 994 photodiodes, 1842 photolithography, in active antenna, 52 photonic analog to digital converters, 1960-70, 1961 avalanche photodiodes in, 1962 channelized, 1964-65, 1964 crosstalk and, 1965 distributed mesh feedback, 1967 electrooptic, 1961-64, 1962, 1963 error diffusion modulator in, 1965, 1965 error diffusion neural networks for, 1966-68, 1967 Gray coding and, 1961, 1962-1963, 1963 history and development of, 1961 Mach-Zehnder interferometers and, 1961-64 mode locked lasers and, 1961 multiple quantum well modulators in, 1967, 1968 Nyquist limit in, 1960 optical cross connects/switches and, 1961 optical demultiplexers in, 1964 optical folding flash type, 1963-64, 1963 oversampling, 1965-68, 1965 oversampling/undersampling in, 1960, 1961 performance analysis for, 1966, 1966 postprocessor in, 1965-66 sampling in, 1960, 1961 self-electrooptic effect device in, 1967, 1968 signal to noise ratio, 1965 signal to quantization noise ratio in, 1966 spectral characteristics of, 1966, 1966, 1967 spectral noise shaping and, 1960 spur free dynamic range in, 1965 time stretching using dispersive optical elements in, 1968, 1968 trends in, 1968-69 wavelength division multiplexing, 1968 photonic feed antenna arrays, 166 photonic integrated circuits, 1759 photonic integrated receivers, 1838 photonic systems, in optical fiber, 1707 photophone, 1849 photovoltaics, 1000, 1000 PHS, 370 physical coding sublayer, Ethernet, 1507-08 physical layer, 15, 1281, 1282 Ethernet and, 1502, 1506-08 OSI reference model, 539 packet switched networks and, 1910 time division multiple access and, 2586-89 wireless communications, wireless LAN and, 1285 physical medium attachment sublayer, Ethernet, 1508 physical medium dependent sublayer, Ethernet, 1508 physical security, 1645 picocells, cochannel interference, 449, 450 piconets, Bluetooth, 314-315, 508 piezoelectric ceramic transducers (acoustic), 34, 35 piezoelectricity, and surface acoustic wave filters, 2444-45 piggybacking in bandwidth reduction, 232-233, 232, 2608 pilot channels, 349, 409, 413 pilot patterns, orthogonal frequency division multiplexing, 1877, 1877 pilot symbols, 398-414, 401, 409, 2053-54 pilot tones, in signal quality monitoring, subcarrier multiplxing, 2272–73, 2272

PIN detectors, 1001

PIN receiver, optical transceivers, 1832-33, 1832

pitch detectors, in speech coding/synthesis, 2372-73 pitch period, in speech coding/synthesis, 2361 pitch prediction filtering, in speech coding/synthesis, 2344-45, 2345 pitch synchronous innovation CELP, 1304 pits and lands, 570, 1736 planar antenna arrays, 61, 142, 148-149, 149, 150, 1374-75, 1377, 1385-89, 1386, 1387 planar inverted F antenna, 193, 195, 195 planar lightwave circuitsoptical crossconnects, 1703-04 Planck's constant, lasers, 1776, 1777 plane radiation pattern antenna, 142-144, 143 plane waves, in holographic memory/optical storage, 2133 plausible values, in maximum likelihood estimation, 1338 Plotkin bound, Hadamard coding, 932 plug and play, 2194, 2310 PMMA, 1742 Pockels effect, in optical modulators, 1742 Pocklington's theorem of cryptography, 615 Poincare arc in optical fiber, 436 point coordination function, in media access control, 0, 1348 point doubling cryptography, 610 point estimation in channel modeling, estimation, tracking, 398 point matching or colocation method, in antenna modeling, 174 point of presence IP networks, 268 point to multipoint communications, 95 point to point communications, 539 carrier sense multiple access and, 339 multicasting and, 1530-31, 1530 point to point protocol, 1644, 2117 point to point tunneling protocol, 1651, 2808 pointers, 638-639, 2498, 2503-08, 2506-2508 Poisson arrival process, in traffic engineering, 488-491, **491**, 497–499, **498** polar cap absorption, in high frequency communications, 949 polarization, 1825, 2065 antenna, 142, 180, 186, 196 dispersion (see polarization mode dispersion), 1970 high frequency communications and, 949 microstrip/microstrip patch antenna and (linear, circular), 1363-64, 1363, 1364 millimeter wave propagation and, 1439-40 multibeam phased arrays and, 1514 optical communications systems and, 1484-85, 1491-93, 1711 optical filters and, 1732 optical modulators and, 1742 optical signal regeneration and, 1759 parabolic and reflector antenna and, 1923 solitons and, 1768 waveguide and, 1390, 1416-17, 1417 wavelength division multiplexing and, 2869 polarization beam splitter, 707, 2272 polarization dependent loss, 1493, 1784, 1843 polarization dispersion, 1711 polarization diversity antenna, 190, 191 polarization efficiency, antenna, 186 polarization fading, 2065 polarization maintaining fiber, 1972 polarization mode dispersion analysis of, 1973, 1973, 1973 bit error rate and, 1971, 1971, 1972, 1973 compensators for, 1970-75 decision feedback equalizers vs., 1973-74, 1973 differential group delay in, 1970-71, 1971 electrical method for, 1973-74, 1973 feedforward equalizers vs., 1973-74, 1973 forward error correction vs., 1971-72 high order compensation for, 1972-73 intersymbol interference and, 1970 least mean square algorithm in, 1974

maximum likelihood sequence estimation and, 1974

pipeline bending using underwater acoustic modem, 20,

20

pipes, 539

optical communications systems and, 1492-93 optical compensation for, 1972, 1972, 1974 optical cross connects/switches and, 1784 optical fiber and, 436, 436, 1711, 1843, 1845 optical transceivers and, 1825 polarization maintaining fiber and, 1972 principal states of polarization in, 1970-71, 1971 self-phase modulation and, 1974 solitons and, 1768, 1770 polarization multiplexing solitons, 1771 policing, traffic, 1659-60 policy-based admission control, 115-116, 116, 118 policy control, flow control, traffic management, 1656 policy decision point, 115-116, 1656 policy enforcement point, 115-116, 1656 policy routing, in flow control, traffic management, 1654-55 polling, media access control, 1345-46 polyphase sequences, 1975-82 autocorrelation in, 1975 binary phase shift keying and, 1976 channelization sequences in, 1975 code division multiple access and, 1975, 1976 crosscorrelation in, 1975 direct sequence CDMA and, 1975-82 EOE sequences in, 1980 even autocorrelation in, 1976, 1977 four phase sequences in, 1977-78 Frank sequence in, 1976 Frank-Zadoff-Chu sequence in, 1978 frequency division multiple access and, 1976 generalized Barker sequences in, 1980-1981 generalized chirplike sequence in, 1978 Gold sequences in, 1976 Golomb sequence in, 1977 Hadamard matrices and, 1976 intercell interference and, 1975 Luke sequence in, 1979 M ary phase shift keying and, 1976 maximum magnitude of OCC in, 1979, 1980 multipath interference and, 1975 multiple access interference and, 1976 odd autocorrelation in, 1976, 1977 optimimum or near-optimum EC in, 1977-79 P1, P2, P3, Px sequences in, 1977 Park-Park-Song-Suehiro sequence in, 1979 perfect EAC in, 1976-77, 1978, 1979 q-phase m sequences in, 1979-80 scrambling sequences in, 1975, 1976 signal to noise ratio, 1975 signature sequences in, 1975 Song-Park sequence in, 1979 spread spectrum and, 1975 spreading sequences and, 1975 time division multiple access and, 1975-76 Walsh sequences and, 1976 population inversion lasers, 1776, 1776 populations, in quantization, 2130 port numbers, 541 post processing, 1030-31, 1047-48, 2263-65, 2264 postamble, 545 postdetection integrator, 445 postfiltering, in speech coding/synthesis, adaptive, 2346 postprocessor, photonic analog to digital conversion, 1965-66 POTS splitters, for modems, 1500 power waveguide, 1396-97 power amplifiers, 530-538, 533 power comparison estimation, in acoustic echo cancellation, 10-11 power control additive white Gaussian noise, 1983 admission control and, 121-122 attenuation and, 1982-83 bit energy to interference power spectral density ratio in, 1983 bit error probablity in, 1983 cochannel interference and, 1982 convergence using standard interference function in, digital phase modulation and, 709

power control (continued) distributed constrained, 1985 distributed scheme for, 1984-86 Doppler effect and, 1983 example of, two-user, 1985-86, 1986 fading and, 1983 feedback information accuracy in, 1986 frames, frame erasure rate in, 1983, 1984 frequency reuse and, 1982 future of, 1987-88 global system for mobile and, 913 interference function in, 1985 iterative algorithm in, distributed systems, 1985 loop delay and, 1986 multipath fading and, 1983 multirate services and, 1987 multiuser detection and, 1987-88 Nakagami fading and, 1983 near far effect and, 1982-83, 1983 open-, closed-, and outer-loop systems for, 1986-87 path loss and, 1983 practical issues on, 1986-87 quality of service, 1982, 1983-84 radio resource management and, 1987, 2092 Rayleigh fading and, 1983 real time vs. non real time services and, 1987 Rice fading and, 1983 shadowing and, 1983 signal to interference ratio in, 1983, 1984, 1986 transmitter power control in, 1982-88 trellis coding and, 2636-37 update rate and, 1986 uplink vs. downlink, 1983 wideband CDMA and, 1986 wireless infrared communications and, 2927 wireless multiuser communications systems and, 1606, 1619 power delay profiles, in indoor propagation models, 2017 power density, antenna, 185, 186 power flux density, local multipoint distribution services, 1268 power gain, antenna arrays, 143 power law crack growth method, optical fiber, 439 power law modulation, 138, 138 power management Bluetooth and, 313-314 cdma2000 and, 366, 367 code division multiple access and, 461-462 continuous phase modulation and, 587, 588, 589 minimum shift keying and, 1457 optical fiber systems and, 1847, 1849 shallow water acoustic networks and, 2208, 2209 wireless communications, wireless LAN and, awake, doze states, 1287 wireless multiuser communications systems and, 1604 wireless packet data and, 2985-86 power spectra of digitally modulated signals, 1988-95 amplitude shift keying in, 1988, 1989-91, 1991 autocorrelation and, 1990 continuous phase frequency shift keying and, 1989, 1991 continuous phase modulation in, 1989, 1991–94, **1992, 1994** cyclostationary or periodically stationary process in the wide sense in, 1989-90 linear modulation in, 1988 linearly modulation in, 1989-91, 1991 nonlinear modulation in, 1988 phase shift keying and, 1989-91, 1991 pulse amplitude modulation and, 1988-91, 1991 quadrature amplitude modulation and, 1989 raised cosine pulse in, 1991-92 spectral shaping in, 1990-91 power spectral density acoustic echo cancellation and, 6 digital phase modulation and, 711-712 free space optics and, 1862 minimum shift keying and, 1463-64, 1464, 1474 orthogonal frequency division multiplexing and, 1869-70, 1870, 1873, 1873 power control and, 1983

powerline communications and, 2001 pulse amplitude modulation and, 2023-24, 2024, 2025 pulse position modulation and, 2035-36 quadrature amplitude modulation and, 2045 random number generation and, 2292-93 serially concatenated coding for CPM and, 2187-88 simulation and, 2287, 2291, 2292-93 space-time coding and, 2325 power spectral density function, optical recording, 579 power spectrum, digital filters, 691-692 power splitters, routing and wavelength assignment in ŴDM, 2104 power sum symmetric functions, cyclic coding, 623 power supplies, active antenna, 53, 65-66 power/booster amplifiers, optical fiber, 1710 powerline communications, 1995-2006 access network for, 1997-99, 1997 applications for, 1996-97 attenuation in, 2000, 2001 automatic request repeat and, 2002, 2004 base station for, 1999 bit error rate in, 2004 broadband, 1997 carrier frequency systems and, 1996 CENELEC and, 1995, 1996, 1997, 2002 channel model for, 2000-2001 community antenna TV vs., 1998 connection admission control in, 2004 coupling in, 1998-99 data rates for, 1995 digital subscriber line and, 1995 direct sequence CDMA and, 2003 echo model in, 2000-2001, 2001 electromagnetic compatibility and, 1995-96, 2001 elements of, 1998-99 error detection and correction in, 2002, 2004 fdma and, 2003 forward error correction and, 2002, 2004 frequency allocation in, chimney approach to, 2001 frequency range in, 2000 frequency shift keying in, 1995 gateways in, 1999 grid of electrical service for, 1996, 1996 impedance in, 2000 in home networks using, 1998 inductance in, 2000 interfaces for, 1998-99 last mile alternative using, 1997-98 LLC sublayer and, 2002 local area networks and, 1998 losses in, 2000 MAC sublayer and, 2002, 2003-04 medium access control in, 1995, 2003-04 modulation in, 1995, 2003 narrowband, 1996-97 network structure for, 1998 noise in. 2001-2002. 2002 orthogonal frequency division multiplexing and, 1995, 2001, 2003 performance problems in, 2002-03 power spectral density in, 2001 protocols for, 1998 quality of service, 2003, 2004 radiation limits in, 2001, 2002 repeaters in, 1999, 1999 reservation protocols and, 2003-04, 2004 resistance in, 2000 ripple carrier signals in, 1996 services offered in, 2002 signal to noise ratio, 2004 standards for, 1996-97, 2002 time division multiple access and, 2003 telephone system vs., 1998 topologies for, 1999-2000, 1999 transmission channel characteristics in, 2000 voltages in, 1995, 1996, 1997 wide area networks and, 1997-98 Powernet EIB, 1996 Poynting vector, 53-57, 61, 165 preamble, 545

preamplifier noise, 1325 preamplifiers, optical fiber, 1710 precoded minimum shift keying, 1462-63, 1462 precoding, partial response signals, 1931-32 preconfigured cycle, optical Internet, 2468-69 predictability, chaos, 421 prediction error in adaptive receivers for spread-spectrum system, 101 prediction error or residual, in vector quantization, 2127 prediction techniques in vector quantization, 2127 predictive coding image and video coding and, 1033-34, 1034, 1037–38, **1037** linear, 1261-68 speech synthesis/coding and, 1300 predictive error filters, image compression, 1063 predictive mapping, image and video coding, 1030 predictive vector quantization, 2127 predistorters, RF power amplifers, compensating for nonlinear distortion, 530-538, 533 preemphasis filtering, 821-823 prefix conditions, compression, 633 preimage resistance, cryptography, 612 preprocessing, image and video coding, 1047-48 preprocessors, sigma delta converters, 2243, 2244 presentation control information (PCI), digital versatile disc, 1738 presentation layer, 540, 1911 pretty good privacy, 1651 preventive congestion control, 112 Prim's algorithm, in multicasting, 1532 prime coding, in optical synchronous CDMA systems, 1810 prime fields, in BCH coding, binary, 239 prime focus, of parabolic and reflector antenna, 2082-83, 2082 prime number generation, 607, 614-615 prime power fields, in BCH coding, binary, 240 primitive BCH coding, 244-252, 622 primitive element, 239, 468 primitive polynomials, 240-241, 468 primitives feedback shift registers and, 792 Golay complementary sequences and, 894-895 paging and registration in, 1915 principal components method, in adaptive receivers for spread-spectrum system, 104 principal state method, in constrained coding techniques for data storage, 577-578 principal states of polarization, 1492, 1970-71, 1971 printed circuit board antenna, 1380, 1428-29, 1429 priority feedback queue scheduling, in wavelength division multiplexing, 657 priority index algorithm, medium access control, 1557 priority queue scheduling, wavelength division multiplexing, 656-657 Privacy and Security Research Group, 1647 privacy enhancing technologies, 1646 private key (see symmetric key/private key encryption) private neighbor sets, cdma2000, 366 private network node interface, 113-114, 204, 205, 1635 proactive routing protocols, ad hoc wireless networks, 2886 probability density evolution, serially concatenated coding, 2167 probability density function blind equalizers and, 289 quadrature amplitude modulation and, 2050-52 traffic modeling and, 1667 transform coding and, 2597 probability of deception, authentication coding, 219, 223 processing elements, neural networks, 1676 processing gain, 96, 348-349, 458-461 product codevector quantization, 2126 product coding, 2007-12 additive white Gaussian noise, 2012 codingwords in, 2008 construction of, 2008-10 direct products in, 2008-09, 2009 encoding in, 2007, 2009 generator and parity check matrices for, 2007-08

product coding (continued) Hamming coding and, 2010-11 Hamming distance and, 2008 iterative coding and, 2010-12 Kronecker product in, 2009-2010 linear block coding as, 2007 low density parity check coding and, 2011 message passing in, 2011 multidimensional coding and, 1538-40, 1539 Shannon limit in, 2011, 2012 single parity check coding and, 2007 Tanner graph for, 2011, 2011 turbo coding and, 2011, 2012, 2727-37 vector quantization and, 2127 program and system information protocol, terrestrial digital TV, 2553 programmable gain amplifier, cable modems, 327, 334 projection algorithm, in channel modeling, estimation, tracking, 407 projective geometry coding, 802-807 projective plane construction, in authentication coding, 220-221 PROMETHEUS project intelligent transportation systems, 503 promiscuous mode operation, 1646, 2888 proof test machine, for optical fiber, 437, 437 propagation, 2067 propagation beam method, 1745 propagation delay, 15, 1250-51 propagation factor or path gain factor, 209 propagation models for indoor communications (see indoor propagation models) propagation of radiowaves (see atmospheric radiowave propagation) propagation of sound, 29-30, 30 attenuation and, 30, 30 density of media vs., 30-31, 31 gas vs. liquid media and, 30-31, 31 Ohm's law analogy to, 31, 31 particle displacement and particle velocity in, 31-32 sound pressure and, 31 sound pressure level and, 32 speed of sound and, 30 wavelength of sound and, 30, 30 propagation path loss, 781-782 propagation time, 1343 proportional integrator (PI), in flow control, traffic management, 1661 PROSAT satellite communications, 198 protection coding, unequal error protection coding, 2762-69 protection cycles, optical Internet, 2469 protection of links or nodes, 1634 protection ratios, cell planning in wireless networks, 379 protection switching, multiprotocol label switching, 1600 PROTECTOR project intelligent transportation systems, 503 protocol data units, cdma2000, 364-365 protocol independent multicast sparse mode, 1534-35 protocol stacks, 541 protocol suites, 541 protocol threading, media access control, 0, 1348 protocols, 538-556 free space optics and, 1851 packet switched networks and, 1910-12 powerline communications and, 1998 satellite communications and, 2113 shallow water acoustic networks and, 2206 prototype waveform interpolative coding, in speech coding/synthesis, 2351 provider edge, 1599 provisioning (see quality of service) proxies, session initiation protocol, 2197, 2198, 2201 proximity coupled microstrip feed line, 1362-63, 1363 proximity detectors, active antenna, 65 proximity effects, in antenna for mobile communications, 189 pseudo multilevel or polybinary signals, in optical receivers, 1825

pseudonoise equalizers, 85

pseudonoise coding cdma2000 and, 362 code division multiple access and, 459 feedback shift registers and, 789 random number generation and, 2293 ultrawideband radio and, 2754 pseudoorthogonal coding CDMA systems, 1809-10 pseudorandom bit sequence, 328 pseudorandom noise carrier sense multiple access and, 349 community antenna TV and, 526 IS95 cellular telephone standard and, 349, 350-351, 354 pseudotraining symbols, spatiotemporal signal processing, 2338 public address systems, acoustic echo cancellation, 6 public key (see asymmetric key/public key) public key infrastructure, 614 public land mobile network, 308 public switched telephone network cellular telephony and, 1479 H.324 standard for, 918-929, 919 IP telephony and, 1172-82, 1173 modems and, 1495 satellite communications and, 877, 2111 software radio and, 2305 speech coding/synthesis and, standards for, 2355 wireless extension to, 308, 308 wireless IP telephony and, 2931-41 pulse amplitude modulation, 2021-30, 2022 additive white Gaussian noise, 2024-25, 2030 amplitude shift keying and, 2022-23 autocorrelation in, 2023-24 automatic gain control and, 2026 bandpass, lowpass, baseband frequencies in, 2022 bandwidth and, 2023 carrier phase recovery in, 2027-28 carrierless amplitude phase modulation and, 336-339 complex envelope in, 2022 correlators in, 2026 Costas loop in, 2028, 2028 cyclostationary processes in, 2023-24 demodulation in, 2024-30 detection of, 2024-30 early late gate synchronizer for, 2029, 2029 energy of, 2022, 2027 error detection and correction in, 2022, 2024-25, 2027 Ethernet and, 1508 Euclidean distance in, 2025 filtering in, 2026, 2029 frequency range for, 2022 Gray coding and, 2027 Gray mapping in, 2023, 2023 likelihood function in, 2026 M ary receiver for, 2025, 2025 matched filters for, 2026, 2029, 2030 maximum a posteriori detectors and, 2026 maximum likelihood detectors in, 2026 memoryless modulation in, 2024 modem and, 2022 modulation coding in, 2024 partial response signals and, 1928, 1933 phase coherent detection in, 2026 phase locked loops in, 2027, 2027 power spectra of digitally modulated signals and, 1988, 1989-91, 1991 power spectral density in, 2023-24, 2024, 2025 pulse position modulation and, 2031, 2034, 2034, 2041 raised cosine pulse in, 2024, 2024 rectangular signal pulse in, 2023 signal representation for, 2022-23 signal space and, 2025 signal to noise ratio, 2026 spectral characteristics of, 2023-24 spectral shaping in, 2024 squaring loop in, 2028, 2028, 2028 symbol synchronization in, 2028-30, 2029, 2028 transmitters and receivers for, 2022-30, 2022 trellis coded modulation and, 2625

voltage controlled clock and, 2029 voltage controlled oscillator and, 2027 pulse amplitude modulation, 7, 1335 pulse amplituide modulation digital magnetic recording channel and, 1323 pulse carver optical transceivers and, 1828 pulse code modulation waveform coding and, 2834-35, 2834, 2834 pulse coding modulation adaptive differential in, 2343, 2354, 2355, 2372, 2382 adaptive differential in, 2820-22, 2822, 2820 companded, in speech coding, 2342 companders and, 527-530, 528, 527 differential in, 2342-43, 2343 image compression and, 1063 linear predictive coding and, 1264 magnetic storage and, 1319 modems and, 1497 permutation coding and, 1954 satellite communications and, 880 speech coding/synthesis and, 1299, 2341-42, 2371, 2372 pulse duration modulation, 2031 pulse interval modulation, 2032, 2032 pulse position modulation, 2030-42, 2031 additive white Gaussian noise, 2037 analog spectrum for, 2036 analog, with nonuniform sampling in, 2033, 2033 baseband signals in, 2034-35 bit error probability in, 2039, 2039 capacity of, 2039-2040, 2040 chaotic, 422, 427-428, 427 chirp modulation and, 441, 444 correlation function in, 2035 cross correlation in, 2042 cyclostationary processes in, 2035 demodulation in, 2036-39 digital signal generation in, 2033-34, 2034 digital spectrum for, 2036 error detection and correction in, 2036-39 error probability evaluation in, 2038-39, 2039 frequency modulation and, 2033 generation and models of signals for, 2032–34, **2032** information rates in, 2039–2040, **2040** Kronecker product in, 2038 local area networks and, 2041 mean and covariance in, 2042 nonuniform sampling in, 2033, 2033, 2041 optical synchronous CDMA systems and, 1809, 1813, 1815–16, **1816** optimal detection of, 2037-38 permutation coding and, 1954 power spectral density in, 2035-36 PSD expression in, proof of, 2041-42 pulse amplitude modulation and, 2031, 2034, 2034, 2041 pulse characteristics in, 2030 pulse duration modulation and, 2031 pulse interval modulation and, 2032, 2032 pulse shape in, 2031 pulse width modulation and, 2031 sampling in, 2030, 2033 serrasoid technique in, 2033 signal to noise ratio, 2039 spectral analysis of, 2034-36 spectral lines (Lebesgue decomposition) in, 2035, 2037, 2038 spectral shaping in, 2032-33 standards for, 2041 synchonism in, 2031 transmitters and receivers in, 2031-32 variations of, 2031-32 pulse regenerator, chaotic, 427-428, 427 pulse shaping, in IS95 cellular telephone standard, 350, 354 pulse width modulation, 2031 pulson application demonstration, ultrawideband radio, 2758 pump lasers, 1778, 1781 pump waves, optical fiber, 1712

punctured convolutional coding, high rate, 979-993 puncturer, in concatenated convolutional coding, 558 pure cycling register, 794, 794, 796-798 pure summing register, 794, 794, 799 push pull operation receivers, 1827 pyramid broadcasting, 236 Q factor (see quality factor) Q.2931 standard ATM, 204 QBone, differentiated services, 674-675 QCELP, speech coding/synthesis, 2354, 2826 q-phase m polyphase sequences, 1979-80 QQ estimator, 1124 Q-switched lasers, 1762, 1762 quad helical antenna, 198 quad loop antenna, 1298 quadratic residue coding, 616-617, 620-621, 933 quadrature amplitude modulation, 715, 1335, 2043-58, 2043. 2179 acoustic telemetry in, 24 additive white Gaussian noise, 2046, 2046 asymmetric DSL and multimedia transmission in, 1576 bandpass in, 2044-45, 2045 bandwidth and, 2043, 2045-46, 2046 bit error probability in, 2043, 2050-52, 2052 bit interleaved coded modulation and, 281 blind carrier recovery in, 2054-56 blind clock recovery in, 2056-57, 2058 blind equalizers and, 292, 296 broadband wireless access and, 319, 320 cable modems and, 324-326, 330-334, 331 carrier recovery or synchronization in, 2052-54 carrierless amplitude phase modulation and, 336-339, **337**, **338** clock recovery (time or symbol synchronization) in, 2052 community antenna TV and, digital video in, 524-527, **526** Costas loop in, 2054, 2054 Cramer-Rao lower bound in, 2055 decision directed carrier recovery in, 2054, 2054 demodulation and detection in, 2047 discrete multitone and, 737 diversity and, error probability and, 2050-52 early late gate synchronizer:, 2057 error detection and correction in, 2047-50 fading and, 2050-52 filtering and, 2046, 2049-50 frequencies for, 2043 Gray coding and, 2043, 2044 high frequency communications and, 954 histogram algorithm and, 2056 home area networks and, 2688 intersymbol interference and, 2045 local multipoint distribution service and, 319, 320 maximal ratio combining and, 2051-52, 2051 microwave and, 2569, 2570 minimum distance algorithm in, 2056 minimum likelihood and, 2054 modems and, 1497, 1498 Nakagami fading in, 2050-52 Nyquist function in, 2049 Nyquist pulses in, 2045, 2046 offset QASK, 2046 optical transceivers and, 1825 orthogonal frequency division multiplexing and, 1868 partial response signals and, 1928 phase locked loop in, 2053-55, 2055 pilot symbols in, 2053-54 power spectra of digitally modulated signals and, 1989 power spectral density in, 2045 predistortion/compensation in RF power amplifiers and, 530, 532, 535, 536 probability density function and, 2050-52 pulse shaping in, 2045 raised cosine pulse in, 2045-46 Rayleigh fading in, 2050-52 receiver for, 2046 shell mapping and, 2221-27, 2221

signal to noise ratio, 2053-55 symbol error probability in, 2043, 2047-49, 2048, 2049, 2050-52, 2050 synchronization and, 2473-85 synchronization in, 2052-57 tapped delay line equalizers and, 1690 terrestrial digital TV and, 2550-55 time recovery with pilot symbols or decision directed in, 2056, **2056** trellis coded modulation and, 2624-35 trellis coding and, 2636-53 turbo trellis coded modulation and, 2738-53 two stage conjugate algorithm in, 2056 in underwater acoustic communications, 41, 43-46 very high speed DSL and, 2791, 2801 voltage controlled oscillator and, 2056 quadrature amplitude shift keying (see quadrature amplitude modulation), 2043 quadrature components, sampling, 2109 quadrature direct digital frequency synthesis, 328, 330, 333 quadrature phase shift keyed, 16, 23, 410, 710, 711, 2179 bit interleaved coded modulation and, 279 broadband wireless access and, 319, 320 cable modems and, 324-326, 330-334, 331 cdma2000 and, 362 continuous phase modulation and, 589 IS95 cellular telephone standard and, 350 local multipoint distribution service and, 318, 319, 320 orthogonal frequency division multiplexing and, 1869, 1945, 1947 peak to average power ratio and, 1945, 1947 predistortion/compensation in RF power amplifiers and, 530, 531 satellite communications and, 881 trellis coded modulation and, 2622-35 trellis coding and, 2637-53 tropospheric scatter communications and, 2693, 2700 turbo coding and, 2704-16 turbo trellis coded modulation and, 2738-53 wideband CDMA and, 2878 wireless multiuser communications systems and, 1610 quadrature spreading, IS95 cellular telephone standard, 354 quadrifilar helical antenna, 197, 197, 199 QUALCOMM, 2112 quality factor, 1478 antenna arrays and, 143, 199 microstrip/microstrip patch antenna and, 1357, 1359-60, 1364 optical fiber and, 1825, 1832, 1846-47 path loss and, 1939-41, 1940 signal quality monitoring and, 2270-71 vector quantization and, 2126 quality of service (see also signal quality monitoring), 549, 1556–58, 1632, 2269 admission control and, 112, 114-117, 116, 120, 121, 122, 126 asymmetric DSL and multimedia transmission in, 1573. 1575-76 ATM and, 204, 205, 207, 266, 272, 273, 550, 552, burst switching networks and, 1804-06 carrier sense multiple access and, 346 cdma2000 and, 359, 363 cell planning in wireless networks and, 372, 379, 379 differentiated services in, 270-271, 668 fiber delay lines and, 1804-06, 1805 flow control and, 1625, 1626, 1653, 1654 general packet radio service and, 866, 868-869 hybrid IntServ-DiffServ in, 271 IMT2000 and, 1099-1101, 1103 integrated services and, 269-270 intelligent transportation systems and, 502 IP networks and, 269-271 IP telephony and, 1172-82, 1173 local multipoint distribution services and, 1269-70

medium access control and, 1558, 1559

mobility portals and, 2192, 2195 multimedia networks and, 1562-68 multimedia over digital subscriber line and, 1571 multiple input/multiple output systems and, 1450 multiprotocol label switching and, 271, 1597-98 neural networks and, 1681 optical cross connects/switches and, 1798, 1804-06 packet rate adaptive mobile receivers and, 1887, 1901 paging and registration in, 1916 power control and, 1982-84 powerline communications and, 2003, 2004 radio resource management and, 2089, 2090, 2094-95 resource reservation protocol in, 270 satellite communications and, 2115, 2117-19 service level agreements and, 270 session initiation protocol and, 2196, 2203 software radio and, 2307 traffic modeling and, 1673 Universal Mobile Telecommunications System and, 387 wireless and, 2915 wireless IP telephony and, 2932-41 wireless packet data and, 2984 wireless sensor networks and, 2995 quantization, 2106 adaptive vector quantization in, 2128 agglomerative methods in, 2128 C means algorithm in, 2129 centroid condition in, 2129 classified vector type, 2127 clustering problems and, 2128 clustering step in, 2129 codebook and, 2123, 2125, 2128-30, 2128 codevector-based approach to training in, 2129 complexity barrier in, 2126 compression and, 639 decoding in, 2125 descent algorithm in, 2129 digital filters and, 686-687 discrete cosine transform in, 2125-26 divisive methods in, 2128 empty cluster problem in, 2129 encoding in, 2125 entropy constrained vector quantization in, 2128 exact vs. approximate methods in, 2126 fine tuner of codebooks in, 2129 finite state vector type, 2127 general optimization methods for, 2130 generalized Lloyd algorithm in, 2128, 2129 genetic algorithms for, 2130 heuristic algorithms in, 2128 image and video coding and, 1026-27, 1030, 1035 image compression and, 1065 iterative methods in, 2128 L stage vector quantization in, 2127 lattice vector quantization in, 2127-28 Llovd's condition and algorithm in, 2125 local optimum in, 2129 losses in, 2123, 2124 lossless compression and, 2123, 2124 lossy compression and, 2123, 2124 mean distance ordered partial search in, 2126 mean removed vector quantization in, 2127 memory requirements of, 2128 nearest neighbor condition in, 2129 nearest neighbor problem in, 2126 nearest neighbor quantization in, 642 open-, closed-, and semi-closed loop, 2127 P median problem in, 2128 pairwise nearest neighbor in, 2128-30 palette generation problem in, 2128 partial distortion search in, 2126 partition-based approach to training in, 2129 populations and individuals in, 2130 prediction error or residual in, 2127 prediction techniques in, 2127 predictive vector quantization in, 2127 product codevector in, 2126 product coding and, 2127 quality and resolution in, 2126

quantization (continued) representative vector in, 2123 reproduction values or points in, 2123 residual or multistage vector quantization in, 2127 residual, in speech coding, 2345-46 robustness in, 2128 running time in, 2128 scalar (see also scalar quantization), 641–642, 1035, 2122–32, 2833 self-organizing maps in, 2130 shape-gain vector quantization in, 2127 sigma delta converters and, 2227-47, 2228 simulated annealing in, 2130 speech coding/synthesis and, 2340-41 splitting method in, 2129 stochastic relaxation in, 2130 stopping condition in, 2129 subvectors in, 2127 tabu lists in, 2130 temperature and cooling schedule in, 2130 training methods and, 2129 training sets in, 2125 transform coding and, 2594, 2597 transforms in, 2125-26 tree structured search in, 2126 tree structured vector quantization in, 2129 triangular inequality elimination in, 2126 uniform vs. nonuniform, 2124, 2124 variance and, 2129 vector (see also vector quantization), 642-644, 1030, 1035–37, **1036**, 1065, 2122–32, 2350, 2372, 2833 - 34Walsh-Hadamard transform in, 2126 Ward's method in, 2129-30 waveform coding and, 2830, 2832-34 quantization error, sampling, 2106 quantizers, transform coding, 2596-97 quantum cascade lasers, in free space optics, 1853 quantum computation, cryptography vs., 615-616 quantum efficiency, 995, 995, 1842 quantum limit, in optical transceivers, 1833, 1837 quarter wave antenna, 193 quartz transducers (acoustic), 34 quasicyclic coding, 2583 quasidynamic mode, in cell planning in wireless networks, 388, 389-390 quasi-Fermi levels, in lasers, 1777 quasiharmonic broadcasting, 236 quasioptic active antenna, 53 quasiorthogonal functions, cdma2000, 362 quasiprime coding, optical synchronous CDMA systems, 1811 quasistatistic approximation, active antenna, 54 queue partitioning, multimedia networks, 1565 queues, 201, 1626, 1627, 1661 queuing delay, in flow control, traffic management, 1653 queuing priority, admission control, 124-125, 125 queuing probability, in traffic engineering, 496 quick look-in decoding of convolutional coding, 2160-61 Rabin encryption, 611-612 radar, 208 active antenna and, 51 antenna and, 169 chaotic systems and, 428-431 clutter in, 429-430, 429 intelligent transportation systems and, 505 ultrawideband radio and, 2761 radial basis function, 102, 1678 radiating near field (Fresnel) region, antenna, 181-182, 182 radiating slot transition, waveguide, 1400, 1400 radiation density, antenna, 185 radiation efficiency, antenna arrays, 143, 184-186 radiation emissions, in powerline communications, 2001, 2002 radiation intensity, antenna, 142-143, 185 radiation patterns, 142-144, 143, 169, 175, 180-181, 181.184

antenna arrays and, 160

antenna for mobile communications and, 190, 192, 193 dipoles, 1257-58, 1258 helical and spiral antenna, 935-946, 936-945 horn antenna and, 1006-17, 1006-16 leaky wave antenna and, 1235, 1239, 1240-41 linear antenna and, 1257-58, 1258 linear antenna and, 1259, 1259 loop antenna and, 1292 microstrip/microstrip patch antenna and, 1357-59, 1358 millimeter wave antenna and, 1425 parabolic and reflector antenna and, 1922-23, 1922, 2080-81, 2081 television and FM broadcasting antenna, 2517-36 transducers (acoustic) and, 32-33, 33 waveguide and, 1417-21, 1418-22 radiation resistance, antenna, 184 radiators, antenna, 180, 199 radio, software, 2304-24 radio access ports, 2088-89, 2091-93 radio astronomy, using parabolic and reflector antenna, 1927 radio frequency components in microelectromechanical systems, 2, 1350 radio frequency interference cable modems, 332 radio link control, in wireless packet data, 2982, 2984 radio link protocol, cdma2000, 359 radio network planning tools, 372, 376, 377 radio refractivity, 2559 radio relay systems, millimeter wave propagation, 1434 radio resource management, 2088-97, 2089 admission control in, 2093-94 automatic response repeat and, 2093 best effort service and, 2094-95 capacity in, 2090 change and, 2093-94 channel to interference ratio in, 2091-93 code division multiple access and, 2090, 2091-93 current approaches to, 2091-93 direct sequence CDMA in, 2090, 2091-93 diversity in, 2093 dynamic channel allocation in, 2091-93 frequency division multiple access and, 2090, 2091–93 frequency hopping and, 2092 general packet radio service and, 873-874 global system for mobile and, 914, 2089 handoffs and, 2093 link gains and, 2089 load sharing in, 2093 power control and, 1987, 2092 problem formation and process of, 2089-91 quality of service and, 2089, 2090, 2094–95 radio access ports and, 2088-89, 2091-93 random channel allocation in, 2091-93 Rayleigh fading and, 2093 resource allocation algorithm in, 2090, 2090 shadowing and, 2093 signal to interference ratio and, 2090, 2091-93, 2092 soft and safe admission control in, 2094 TCP/IP and, 2094-95 time division multiple access and, 2090, 2091-93 radioastronomy, waveguide, 1392 radiolocation (see also wireless, location in), 2959 radiowaves, propagation of (see atmospheric radiowave propagation) rain attenuation, 215-216 microwave and, 2560 millimeter wave propagation and, 1270-72, 1271, 1440-45, **1440**, **1441** RAINBOW, 1720 raised cosine modulation, 585 raised cosine pulse power spectra of digitally modulated signals and, 1991-92 pulse amplitude modulation and, 2024, 2024 quadrature amplitude modulation and, 2045-46 raised cosine spectrum, partial response signals, 1929-30, 1930 RAKE processing/RAKE receivers, 2481

adaptive receivers for spread-spectrum system and, 108 chann/in channel modeling, estimation, tracking, 411 diversity and, 732, 734 fading and, 787-788 IS95 cellular telephone standard and, 356 location in wireless systems and, 2968-70, 2969 mobile radio communications and, 1481 multicarrier CDMA and, 1523-24, 1523 orthogonal frequency division multiplexing and, 1878 packet rate adaptive mobile receivers and, 1886, 1887, 1898, 1900, 1901 shallow water acoustic networks and, 2209 software radio and, 2307 ultrawideband radio and, 2757-59 Universal Mobile Telecommunications System and, 387, 388 RAMAC systems, 1320, 1321 Raman amplifiers, 1709, 1842 Raman scattering, 1491, 1684-85, 1685, 1712 random access protocols, traffic engineering, 499-501, 499 random capacity concept, wireless transceivers, multiantenna, 1580-81 random channel allocation, radio resource management, 2091-93 random coding, 2157 random delay, media access control, 1346 random early detection, 1661, 1661, 1627, 1628, 1630 random early marking, flow control, traffic management, 1661 random number generation arbitrary distribution and, 2292 autoregressive moving average in, 2293 binary and nonbinary sequences in, 2293 Box-Mueller method in, 2292 correlated Gaussian sequences in, 2292-93 cryptography and, 607, 614-615 finite impulse response and, 2292-93 Gaussian, 2292 inverse transform method in, 2292, 2292 linear congrential algorithm in, 2292 Marsaglia-Zamann algorithm and, 2292 Monte Carlo simulation and, 2292-93 power spectral density and, 2292-93 pseudonoise sequences and, 2293 simulation and, 2291-93 uniform, 2292 Wichmann-Hill algorithm in, 2292 random phase channels, expectation maximization algoritĥm, 772 random phase mask (RPM), in holographic memory/optical storage, 2133 random signature sequence for CDMA, 2276 random service order, in traffic engineering, 498 random vectors, in maximum likelihood estimation, 1338 random walk, 412, 1918 randomization, in community antenna TV, 526 range, of underwater acoustic communications, 37-38 range dependence, in path loss, 1941 range-based telemetry (see also acoustic telemetry; telemetry), 26–27, **26**, **27** Rao-Wilton-Glisson basis functions in antenna modeling, 176 rate-based control schemes, ATM, 551-552 rate compatible channel coding, in speech coding/synthesis 2355 rate compatible punctured convolution coding, 2355 rate distortion theory additive white Gaussian noise, 2069-80 Bernoulli sources in, 2073 binary symmetric channel in, 2073 Blahut algorithm in, 2075 CEO problem and, 2076 channel coding and, 2069 convergence in, 2075 doubly matched configurations in, 2076 Gray coding and, 2075 Hamming distortion in, 2073 history and development of, 2070-71

rate distortion theory (continued) information transmission inequality in, 2070 Lempel-Ziv coding and, 2076 mean sequence error in, 2069-80 mutual information rate in, 2070 Shannon or channel capacity and, 2069 signal to noise ratio, 2069, 2070 single letter fidelity in, 2072–76 source coding and, 2069 water pouring result in, 2069 rate equations, for lasers, 1778 ray tracing, in indoor propagation models, 2019, 2019 Rayleigh criterion, 212–213, **213**, 1750 Rayleigh fading/distribution, 785-786 antenna for mobile communications and, 190 bit interleaved coded modulation and, 278, 280, 281, 283.285 cellular communications channels and, 394 in channel modeling, estimation, tracking, 410 chirp modulation and, 446 diversity and, 732, 733 expectation maximization algorithm and, 776-778, 778 in underwater acoustic communications, 40 location in wireless systems and, 2967 microwave and, 2563 multiple input/multiple output systems and, 1455-54, 1454 power control and, 1983 quadrature amplitude modulation and, 2050-52 radio resource management and, 2093 satellite communications and, 1226-27, 1226, 1227 simulation and, 2291 wireless and, 2920-22 Rayleigh scatter, 1271, 1855-57 RCE photodectors, 1002-1003, 1003 reachability concept, in multicasting, 1536 reactance, in loop antenna, 1295, 1295 reactive congestion control, 112 reactive near field region, antenna, 181-182, 182 read process CDROM and, 1734 digital magnetic recording channel and, 1323-24 digital versatile disc and, 1737 hard disk drives and, 1320 magnetic storage and, 1320, 1326, 1327-28 optical memories and, 1733 real time control protocol, 1662 real time protocol, 1181, 2934-35 real time service, medium access control, 1555-58 real time streaming protocol, 2438, 2979 real time transport protocol, 2436-37 real time variable bit rate, 206, 267, 551, 1658 rebroadcasting, ALOHA protocol, 128 recall, in neural networks, 1675 received signal phase, in wireless systems, 2690 received signal strength, in wireless systems, 2690 receiver available time table, medium access control, 1554, 1555 receiver oriented earliest available time scheduling, 1554-55 receivers acoustic telemetry in, 23, 23 adaptive receivers for spread-spectrum systems, 95-112 blind multiuser detection and, 304-306, 305 chann/in channel modeling, estimation, tracking, 398-408 companders and, 527-530 continuous phase frequency shift keying and, 594-598, 595 continuous phase modulation and, 591, 591, 592 direct detection, 1825 discrete multitone and, 740, 742 free space optics and, 1851-52, 1852 heterodyne receivers, 1835 holographic memory/optical storage and, 2136–37, 2137 homodyne, 1835 local multipoint distribution services and, 1268 microwave and, 2567-70, 2568

minimum shift keying and, 1462-67, 1462, 1465 multicarrier CDMA and, 1522-25, 1522 multiple input/multiple output systems and, 1450-56, 1450 optical (see also optical transceivers), 1824-40 optical communications systems and, coherent, 1484, 1486-88, 1486, 1487, 1488 optical fiber and, 1709 optical memories and, 1733 optical synchronous CDMA systems and, 1815, 1815 orthogonal frequency division multiplexing and, 1867-71 packet rate adaptive (see packet rate adaptive receivers for mobile communications) packet rate adaptive mobile receivers and, 1887-88, 1887 photonic integrated, 1838 pulse amplitude modulation and, 2022-30 pulse position modulation and, 2031-32 Q factor in, 1832 quadrature amplitude modulation and, 2046 reduced search, for CPM, 592 satellite communications and, 2115 sensitivity in, 1825, 1833-34 sidebands in, 1826 signature sequence for CDMA and, 2275–76, 2275 superheterodyne, 1478 tropospheric scatter communications and, 2699-2703 ultrawideband radio and, 2757, 2757 in underwater acoustic communications, 42-45, 42 wavelength division multiplexing and, 651 wireless infrared communications and, 2926 wireless multiuser communications systems and, 1608 - 20receiving antenna, 1260, 1260 reciprocal multidimensional coding, 1540 reconstruction of images, 1079-94, 1081-92 recordable DVD-R media, 1738 recovery, 1650 rectangular parity check coding, 1543-44 recurrent neural networks, 1680 recursion, in adaptive receivers for spread-spectrum system, 105 recursive least mean squares algorithm, 101 recursive least square algorithm acoustic echo cancellation and, 8-9, 8 in channel modeling, estimation, tracking, 404, 414, 415 equalizers and, 82, 84-85, 85, 90 packet rate adaptive mobile receivers and 1886, 1883, 1887 underw/in underwater acoustic communications, 44 recursive mean squares equalizers, 286 recursive systematic convolutional coding, 556-557, 557, 2182, 2705–07, 2706 Red Book, 1736 reduced state sequence estimation, 81, 1933 reduced-rank adaptive MMSE filtering, 103-104 reduced-rank detection, adaptive receivers for spreadspectrum system, 104-105 redundancy, 1632 image and video coding and, 1027-28 trellis coded modulation and, 2623 redundant array of independent disks, 474-475, 1322 Reed-Muller coding, 628, 929, 932-933, 1950 Reed-Solomon coding (see also cyclic coding), 238, 253-262, 616-630 Berlekamp decoding algorithm for, 624-625 bit error rate in, 473 block error rate in, 473 block missynchronization detection in, 471-472 bounded distance decoding in, 254 cable modems and, 330, 332 CDROM and, 1735 Chien search decoding and, 256-257, 260, 470, 617 community antenna TV and, 526 compact disk and, 626 connection polynomial in, 257-258 constrained coding techniques for data storage and, 576 cyclic coding in, 469 decoding in, 254–261, 469–470, 622–626

deep space telecommunications and, 629 elementary symmetric functions in, 623 encoder for, 254, 254, 468-469 erasure filling decoding in, 259-261 error correcting coding in, 470, 472-474 error detecting coding in, separate vs. embedded, 474 error locators in, 623 error magnitudes or error values in, 254 error rate definitions for, 473 Euclid's algorithm and, 617 extension of, 467 feedback shift register and FSR synthesis in, 257-259 generalized minimum distance decoding in, 261 generating functions in, 623 hard decision decoding algorithms for, 475 hardware vs. firmware implementation of, 470-471 interleaving vs. noninterleaving in, 472 large sector size and, 475 linear coding in, 469 locator fields in, 253 magnetic recording systems and, 2249 magnetic storage and, 1326 Massey-Berlekamp decoding algorithm for, 257-259, 260, 470, 617, 625-626 maximum coding and, 254 maximum distance separable coding and, 254 modified syndromes for decoding in, 259-260, 469-470 multiple input/multiple output systems and, 1456 Newton's identities and, 255, 257, 623, 625 performance and 472-474 Peterson's direct solution method for, 255-257, 260, 617 polynomials and, 254 power sum symmetric functions in, 623 primitive elements in, 468 primitive polynomials in, 468 primitive vs. nonprimitive types, 253 redundant array of independent disks and, 474-475 Reed-Solomon coding for magnetic recording channels and, 467-475 sequential decoding of convolutional coding and, 2158 serially concatenated coding and, 2164 soft bit error rate in, 474 soft decision decoding algorithms for, 261, 475 symbol error rate in, 473 symbol fields in, 253 syndrome equations for decoding in, 255, 623 systematic coding in, 469 t error correcting coding and, 253 tape drive ECC and, 474 trellis coding and, 2640 turbo coding and, 2703 in underwater acoustic communications, 43 very high speed DSL and, 2800 reference signals, in equalizers, 85-86 reflectarray microstrip/microstrip patch antenna, 1387, 1387 reflection, 2065 cellular communications channels and, 393 indoor propagation models and, 2013, 2018 microwave and, 2556-57, 2557 parabolic and reflector antenna and, 2082, 2086 satellite communications and, 196 ultrawideband radio and, 2759-60 reflection coefficient, 3, 1401, 1402 reflector antenna (see also parabolic and reflector antennas), 169, 179, 180, 184, 187, 1006-17, 1006-16, 1425-26, 1426, 1427, 1425, 2080-88 refracted near field method, in optical fiber, 435 refraction, 210-211, 210 microwave and, 2558-60, 2559 millimeter wave propagation and, 1434-36, 1435, 1445 refractive index lasers and, 1779 millimeter wave propagation and, 1434-36, 1435, 1445 optical fiber and, 1686, 1715, 1765 optical modulators and, 1745, 1745 solitons and, 1764

regeneration, 1319, 1759-64 regenerators, optical fiber, 1707 region 2 skywave method, 2061–62 registration, paging, in mobile networks, 1914-28 regular pulse excitation algorithm, 2824 regular pulse excitation with long term predictor, 1304, 2356 relative spectral method, 2378 relaxed CELP, 2827 relaxed linear time invariant systems, 689-90 reliability, 1631-44, 2067 asynchronous transfer mode and, 1633-35 automatic repeat request and, 1632 backup schemes and, 1634-35 broadband and, 2655 circuit switched networks and, 1632 compression and, 631 cycle covers and, 1638-39, 1638 cyclic redundancy check in, 1633 dynamic restoration in, 1635 fail stops in, 1632 failure and fault detection/recovery in, 1631-34 fault isolation boundaries in, 1632 fiber distributed data interface and, 1637 free space optics and, 1865 high speed/Gigabit LANs and, 1640-42, 1642 intermittent failures and, 1631 link and node-based schemes for, 1635 link rerouting in, 1633-34, 1634 Menger's theorem and, 1635 mesh networks and, 1637-39, 1638, 1639 metropolitan area networks and, 1632 minimum spanning tree in, 1639-40 models for, 1632 multimedia networks and, 1562 multiprotocol label switching and, 1640 optical fiber and, 439, 1636, 1636 optical modulators and, 1746 packet switched networks and, 1632, 1639-40 path and link monitoring in, 1633 path-based schemes for, 1634-35, 1634 protection of links or nodes in, 1634 quality of service and, 1632 redundancy and, 1632 rings for, 1635–37, **1636** self-healing rings in, 1635, 1637, 1638 SONET and, 1634, 1635 subnetwork connection protection and, 1635 topologies for, 1632-33 transmission control protocol and, 1632, 1640 transport protocols for optical networks and, 2615-16 wide area networks and, 1632 wireless multiuser communications systems and, 1605 reliable protocols, 543 remote defect indicator, ATM, 207 remote imaging, in underwater acoustic modem, 20-21, 21 remote method invocation, distributed intelligent networks, 725, 727 remote sensing, parabolic and reflector antenna, 1928 remotely operated vehicles, acoustic telemetry, 28 renewal models, in traffic modeling, 1666, 1667 repeaters, 1504-05, 1999, 1999 replica in SPC coding, 1541-42 reply storms, ad hoc wireless networks, 2888 reproduction codebook, transform coding, 2596 reproduction values or points, quantization, 2123 request to send, 346 Research and Development in Advanced Communication Technology in Europe, 397 reservation protocol, 1558 reservation ALOHA, 130 reservation-based protocols burst switching networks and, 1801-02 flow control, traffic management and, 1655, 1656-57 media access control and, 1343, 1347-48, 1552-54 mobility portals and, 2195 multimedia networks and, 1567 optical cross connects/switches and, 1800, 1801 powerline communications and, 2003-04, 2004

satellite communications and, 1232 statistical multiplexing and, 2420-32 virtual private networks and, 2809 residential broadband, 2666-73 residual echo suppressing filter, 1-7, 2 residual or multistage vector quantization, 2127 residual quantization, in speech coding/synthesis, 2345 - 46resilient packet ring, 1637 resistance antenna, 184 loop antenna and, 1295 powerline communications and, 2000 resistive impedance, in active antenna, 49 resolution, 1923, 2126 resolvers, 548 resonance, 33-34, 48 resonant frequency/resonant dimension lasers and, 1779 microstrip/microstrip patch antenna and, 1359, 1359, 1361, 1361 resonant scatter, in free space optics, 1855-57 resonators microelectromechanical systems and, 1354-55 surface acoustic wave filters and, 2454-57, 2455 resource allocation ATM, 552 flow control, traffic management and, 1653, 1663, 1663 IP telephony and, 1180 multimedia networks and, 1563, 1567 packet switched networks and, 1908 power control and, 1987 radio resource management and, 2088-97, 2089 wireless local loop and, 2951 resource allocation algorithm, 2090, 2090 resource-based admission control, 112, 118 resource reservation protocol admission control and, 114-115, 115, 116 flow control, traffic management and, 1655, 1656-57, 1657, 1659 IP networks and, 270 mobility portals and, 2195 multimedia networks and, 1567, 1568 statistical multiplexing and, 2420-32 resource reservation protocol for tunneling, 1596-97 response time, in satellite onboard processing, 482 retraining modems, 1498 retrograde orbits, 1248 retroreflection, in active antenna, 65 retroreflectors, in diffraction gratings, 1755 return loss, parabolic and reflector antenna, 1924 return to zero optical signal regeneration and, 1759-1763 partial response signals and, 1933-34, 1933 wireless infrared communications and, 2927, 2928 return to zero DPSK, 1825 return to zero on off keying, 1827-28, 1827 reverberation time, in acoustic echo cancellation, 2 reverse link, 349-357, 349, 353-355, 355, 362-363, 363, 364, 367 reverse shortest path tree, multicasting, 1534 reversible variable length codes, 2977, 2977 RF power amplifiers adjacent channel interference in, 530 AM/AM characteristics in, 531, 531 AM/PM characteristics in, 531, 531 amplitude/phase predistorter for, 533, **533** binary phase shift keying in, 531 binary PSK and, 530 bit error rate in, 530, 535-536, 535, 536 compensation methods for nonlinear distortion in, 532-537 direct vs. indirect architecture and learning in, 535 inband interference in, 530 interference in, 530 inverse fast Fourier transform in, 532 learning architecture of Volterra-based predistorter in, 535, 535 lookup table for predistortion in, 533-534 minimum mean square error predistortion in, 532–533, **532**

modulation and, 530 nonlinear characteristics of, 531 nonlinear distortion in, 530-538, 533 orthogonal frequency division multiplexing and, 530, 531-532, 535 peak to average power ratios in, 530, 532 phase shift keying and, 530 predistorters for, compensating for nonlinear distortion, 530-538, 533 quadrature amplitude modulation and, 530, 532, 535, 536 quadrature phase shift keying and, 530, 531 saturation and, 530 sensitivity of OFDM systems to nonlinear distortion in, 531-532 signal to noise ratio in, 530 simulation experiments in nonlinear distortion in, 535-537 solid state power amplifiers, 531, 535 traveling wave tube amplifier, 531, 532, 535 Volterra-based predistorer for, 533, 534-535, 534 rhomboid antenna, 180 Riblet linear antenna arrays, 147, 147 Rice fading, 785-786 cellular communications channels and, 394 power control and, 1983 satellite communications and, 1226-27, 1226 simulation and, 2291 wireless and, 2919 ridged waveguide transition, 1400, 1400 Rimoldi's transmitter, in minimum shift keying, 1467-68, 1468 ring cover/node cover, in optical Internet, 2468 ring linear Golay coding, 890-891 ring modulator, in amplitude modulation, 139, 139 ring topologies backhauling in, 1636, 1636 dense WDM and, 748-757, 749 loopback in, 1636, 1636 matched nodes in, 1637, 1637 optical fiber and, 1716, 1716 reliability and fault tolerance in, 1635-37, 1636 SONET and, 2495-96, 2496 ripple carrier signals (RCS), in powerline communications, 1996 RLC networks and active antenna, 64 rlogin, 2608 roaming, 915-916, 916, 1287 robust header compression, 2936-37 robustness, in quantization, 2128 Rochelle salt transducers (acoustic), 34 role-based access control, 1649 rollover effect, in speech coding/synthesis, 2366 rooftop basis functions, in antenna modeling, 176 root matching, in antenna arrays, synthesis, 154 root mean square value, in parabolic and reflector antenna, of surface, 1924 rotational invariance, in trellis coded modulation, 2632-33 roughness factors (specular effects), in radiowave propagation, 211-213, 212, 213 round trip time flow control and, 1627, 1672 transmission control protocol and, 554, 2609, 2612 rounding (see quantization) routers and routing, 541, 547-550 ad hoc wireless networks and, 2886-87 ATM and 205 Bellman-Ford algorithm in, 2208 burst switching networks and, 1802-04 constraint based, 1654-55 core routers in, 1802, 1804, 1804 Dijkstra algorithm and, 2208 distributed intelligent networks and, 719-29, 722, 726 distributed, 1566 edge routers in, 1802 egress edge router in, 1803, 1803 general packet radio service and, 869-870, 873 global system for mobile and, 914-915, 915 hierarchical, 1566

3056 INDEX

routers and routing (continued) ingress edge routers in, 1802, 1803 multicasting and, 1531, 1531, 1533, 1566 multimedia networks and, 1563, 1566 multiprotocol label switching and, 1590-1601 optical Internet and, 2469 optical multiplexing and demultiplexing and, 1749 packet switched networks and, 1907, 1909–10, 1913 satellite communications and, 2115, 2118 satellite communications and, multihop satellite routing in, 1254 session initiation protocol and, 2200-01 shallow water acoustic networks and, 2208, 2211 source, 1566 virtual private networks and, 2809 wavelength division multiplexing and, 2097-2105, 2839-40, 2840, 2864 wavelength routing networks in, 1798, 1799-1800 wireless IP telephony and, 2939, 2939 wireless sensor networks and, 2994 routing and wavelength assignment, 1800, 2845, 2097-2105 adaptive routing and, 2102 call blocking probability in, 2103 constraints on RWA in, 2098-99 dynamic routing and wavelength assignment in, 2101-04 electrooptic bottlenecks in, 2098 embedded topologies in, 2101 fairness in, 2103 fanout of power splitters in, 2104 first fit routing in, 2102 fixed-alternate routing and, 2102 least loaded routing in, 2102 light trees and, 2100 light trees in, 2104 lightpaths (gamma channels) in, 2098, 2101 logical or virtual topologies for, 2100-01 maximum reuse routing and, 2102-03 minimum reuse routing and, 2103 mixed integer linear programming in, 2100, 2101 mixed integer programming in, 2100 multicast capable OXCs, 2104 multicast routing and wavelength assignment in, 2104-05 optical cross connects/switches in, 2098-2105, 2099 path mapping in, 2101 pathlength and, 2102 power splitters in, 2104 sparse light splitting in, 2105 static routing and wavelength assignment in, 2100-01 switching in, 2098 topologies for, 2100-01 unfairness factor in, 2103 wavelength converters in, full, limited, fixed, 2099-2100, 2099 wide area networks and, 2098 routing control security, 1649 routing information protocol, 549, 1534 routing tables, 269, 549 Rowland circle, diffraction gratings, 1751, 1752 RS-232 modems, 1495 RSA (Rivest-Shamir-Adleman) algorithm, 335, 606, 611, 615, 1152, 1156, 1649 RS-xxx interfaces, in acoustic modems for underwater communications, 18 RTP control protocol, 2438 RTS/CTS flow control modems, 1497 RTSP, 2198 rubber duck, 193 Rudin-Shapiro construction, 893, 1951-52 run length coding, 1030, 1046 run length limited, 579-581 constrained coding techniques for data storage and, 571–573, **571** magnetic recording systems and, 2248, 2249, 2254 magnetic storage and, 1327 partial response signals and, 1934 Runge-Kutta integration, in chaotic systems, 422, 424 running digital sum, in optical recording, 579 running time in quantization, 2128

SAFER+ Bluetooth, 316 Sagnac interferometers, 1749 Salutation mobility portals, 2194 sample matrix inversion, 1886-1903 sampling, 2106-11 analog signal, 2106-11 analog to digital conversion and, 2106-11, 2106 antenna arrays and, synthesis, 154 bandpass signals and, 2108–2111, **2109** blind equalizers and, 286, 287 coding and, 2106 community antenna TV and, 522 compression and, 631-632 cutoff frequency and, 2110 frequency for, 2106 frequency-domain relationships in, 2107-08, 2108 fundamental range in, 2107 holographic memory/optical storage and, 2138 image and video coding and, 1026 image sampling and reconstruction and, 1079-94, 1081-92 interval for, 2106, 2107 inverse Fourier transform and, 2107 lowpass signals and, 2109-10, 2109 Nyquist sampling rate in, 2107 photonic analog to digital conversion and, 1960, 1961 pulse position modulation and, 2030, 2033 quadrature components in, 2109 quantization and, 2106, 2122-32 quantization error and, 2106 simulation and, 2294, 2294 speech coding/synthesis and, 2370, 2340-41 theorem of, 2108 time-domain relationships in, 2108, 2108 waveform coding and, 2830–32, 2831, 2837 sampling frequency, 2106 sampling interval, 2106, **2107** sampling rate, simulation, 2287 sampling theorem, 2108 sand and dust attenuation, millimeter wave propagation, 1443 satellite communications (see also communications satellite onboard processing; land mobile satellite communications), 208, 876-885, 1223-24, 2179, 2653 access points (AP) and, 2117 ACES, 196 acoustic telemetry in, 22 adaptive differential PCM and, 880 additive white Gaussian noise and, 1251 Advanced Communications Technology Satellite in, 1227, 1228 advanced mobile ... service and, 2116 ALOHA protocols in, 1232, 1253 American Mobile Satellite Corporation in, 2112 AMSC, 196 antenna arrays and, 141, 169, 189, 196-199, 877-878 antenna direction and, 1223, 1228-29 apogee and perigee in orbit of, 1248 ascending node in orbit of, 1248 astra return channel system and, 2120 asynchronous DSL and, 2121 asynchronous transfer mode and, 2113, 2115, 2120 AUSSAT, 196 automatic repeat request and, 224-231, 879 base station location and, 2117 beam patterns for, 877-878 Big Leo systems in, 1251 binary frequency shift keying in, 1225, 1225 binary phase shift keying in, 1225, 1225, 1230 bit error rate in, 881, 1224, 1225, 1227, 1230, 2120 blind multiuser detection and, 298–307 block coding in, 1229–30, **1230** broadband and, 2112-13, 2113, 2115, 2655, 2656, 2664-66, 2665, 2666, 2671-73 broadcast satellite service in, 877, 1251 broadcasting using, 2112 C band in, 877, 1251, 2113 carrier sense multiple access and, 339-340, 340 cavity backed cross slot antenna in, 197-198 cellular telephony and, 2112

channel characteristics in, 1224-29 circuit switched network architectures and, 1253-54 code division multiple access and, 458, 879, 881, 1231-32, 1231 community antenna TV and, 514 Comsat and, 876 congestion control in, 2120 constellation of satellite in, 1247, 1248–50 continuous phase modulationin, 1225 convolutional coding in, 1229-30, 1230 coverage or footprint in, 1249, 2111 crossed drooping dipole antenna in, 197, 197 delay in, 879, 2112 demand assignment multiple access and, 879 differential phase shift keying in, 1225, 1225 digital subscriber line and, 2121 digital voice and television in, 880 digitial video broading in, 2112, 2671-73, 2672 directional antenna in, 198 diversity techniques in, 1230-31 Doppler shift in, 196, 197 downlinks in, 1223, 1223 ECHO, 196 effective isotropic radiated power in, 881 Ellipso, 196 error detection and correction in, 1223, 1229-31, 1230, 1231 ETS-V, 198 European mobile satellite and, 2112 fading in, 1223, 1226-29, 1226, 1227 feeder links for, 1251 figure of merit in, 1229 fixed satellite services in, 877, 1251 forward error control in, 878, 1223, 1229-31, 1230, 1231 forward/reverse path in, 1223, 1223 free space optics and, 1850 frequencies for, 877, 1223-24, 1224, 1251, 2113 frequency division multiple access (frequency division multiple access and, 829, 878-881, 1231-32, 1231 frequency shift keying in, 1225 future of, 1255-56 gain to system noise in, 196, 1229 gateways for, 881, **882**, 2114 Gaussian minimum shift keying in, 1225, 1225 general packet radio service and, 2117, 2118 generations of systems in, 2112 geostationary satellite in, 196, **196**, 1223, 1224, 1231, 1232, 1248–50, 2113 Global Positioning System, 198, 1224, 1254 global system for mobile and, 2116 Globalstar in, 196, 1231, 1247, 1250, 1250, 1251, 2112 handoffs in, 1252, 1254, 2118-20 highly elliptical orbit satellite in, 1249 IMT2000 and, 2116 information vs. coding rate in, 1229 INMARSAT and, 196, 198, 876, 1224, 1227, 2112 Intelsat and, 876–885, 2112 interference and, 1251 intermediate circular orbit systems in, 1224 Internet and, 2113–15, 2114, 2115, 2120–21 Internet protocol (IP) and, 1253 Internet service providers and, 2115 internetworking units and, 2116, 2117 intersatellite handoffs in, 2119 intersatellite links in, 1224, 1252, 2113 ionospheric scintillation in, 196, 197 IP addressing and, 2117 IP networks and, 268, 2111-22 Iridium and, 196, 1247, 1250, 1250, 1251, 1253, 1253, 2112 iterative coding in, 1229-30, 1230 Ka band in, 877, 1251, 2113 Ku band in, 877, 1251, 2113 L band, 196, 1251 launch of, spacecraft used for, 1251-52 link budgets for, 883-884, 1229 link performance in, 1251 location management and, 2117-18

satellite communications and, 196

satellite communications (see also communications satellite onboard processing; land mobile satellite communications) (continued) location registration in, 1253-54 low earth orbit satellite in, 196, 196, 1223, 1224, 1231, 1247-56, 2112, 2119 Marisat and, 876 maximal ratio combinining in, 1230 media access control and, 879, 1342–49 medium earth orbit satellite in, 196, 196, 1223, 1224, 1231, 1232, 1249, 2112 microelectromechanical systems and, 1349 microstrip patch antenna in, 197, 197 millimeter wave propagation and, 1434 mobile nodes in, 2118 mobile satellite service in, 877, 1251, 2112 mobility management and, 2116-17, 2118, 2119-20 modulation in, 1225, 1225 Molnya orbit in, 1250 MPEG compression and, 880 MSAT, 196, 198 multibeam phased arrays and, 1519 multicarrier frequency division duplex in, 2116 multihop satellite routing in, 1254 multipath fading in, 1226-27, 1226 multipath interference and, 196 multiple access in, 1253 Navigation System with Time and Ranging in, 198 noise in, 1224-25 NSTAR in, 2112 omnidirectional antenna for, 197-198 onboard processing for (see communication satellite onboard processing), 880-881, 2113, 2114 onboard switching in, 2113 Orbcomm in, 1251 orbital geometry for, 196, **196**, 877, 1223, 1224, 1248–49, **1248**, 2112, 2113 OSI reference model and, 2118 packet error rate in, 881 packet switched architectures and, 1255 parabolic and reflector antenna and, 1928 path loss in, 1223, 1225-26, 1225 payloads in, 881-883, 882 phase shift keying in, 1225 point to point protocol and, 2117 polarization in, 196 propagation delay and, 1250-51 propagation in, 1223 PROSAT, 198 protocols for, 2113 public switched telephone network and, 877, 2111 pulse coding modulation and, 880 quad helical antenna in, 198 quadrature phase shift keying and, 881 quadrifilar helical antenna in, 197, 197 QUALCOMM in, 2112 quality of service and, 2115, 2117, 2118-19 Rayleigh channels, Rayleigh fading in, 1226-27, 1226, 1227 receivers for, 2115 reflection and, 196 reservation protocols in, 1232 retrograde orbits and, 1248 Rice factor in, 1226-27, 1226, 1227 routing and, 2115, 2118 satellite diversity in, 1231 satellite used in, 196, 1223, 1224 scattering and, 196 seams in orbits of, 1250 selective acknowledgment in, 2120 shadowing in, 1223, 1227-28 short backfire antenna in, 198, 198 signal to noise ratio in, 1224, 1230 single channel per carrier in, 878 slant range in, 1225-26, 1226 space communications protocol standards in, 1233 spatial diversity in, 1230-31 split TCP in, 2120 spoofing in, 2120 spot beams in, 877-878, 1249 standards for, 2113

station keeping in, 1248 subscriber links for, 1251 switching in, 2113 target probability in, 2119-20 TCP for transactions in, 2120 TCP/IP and, 2113, 2120 technical issues for implementation of, 2116-19 Teledesic, 196 third-generation wireless systems and, 2115-16, 2115 time division multiple access (time division multiple access and, 878-881, 1231-32, 1231, 1253 tracking in, 1252 transmission control protocol and, 2120 transmitters for, 878, 2111 transponder for, 878 trellis coded modulation and, 2631 Tundra orbit in, 1250 turbo coding and, 1229-30, 1230 UMTS terrestrial access radio network and, 2116 universal mobile telecommunication service and, 2116 uplinks/downlinks in, 877, 1223, 1223, 2115 Van Allen belts and, 2112 very small aperture terminal in, 879-880, 1247 voice over IP and, 2121 Walker delta or rosette constellatioin in, 1250, 1250 Walker star or polar constellations in, 1250, 1250 waveguide and, 1391-92, 1392 wideband CDMA, 2116 wireless LAN and, 2118 satellite digital audio radio service, 680 satellite diversity, 1231 satellite transport protocol, 1233 Sato algorithm, 92, 291 saturation, in predistortion/compensation in RF power amplifiers, 530 saturation recording, magnetic recording systems, 2248-49, 2249 scalability flow control and, 1626 optical filters and, 1732 signal quality monitoring and, 2269 switches, ATM and, 201 wavelength division multiplexing and, 2865 scalar potentials, in loop antenna, 1290-91, 1294 scalar quantization, 2122-32 compression and, 641-642 distortion rate function in, 2123 fixed rate coding and, 2123 Huffman coding and, 2124 image and video coding and, 1035 Lloyd's condition and algorithm in, 2125 transform coding and, 2601-02, 2602 uniform vs. nonuniform, 2124, 2124 variable length coding in, 2123 waveform coding and, 2833 scaling mechanisms, in multimedia networks, 1563, 1566 scan blindness (see blindness in microstrip antenna) scan loss, multibeam phased arrays, 1517 scan mode, Bluetooth, 311-312 scanning image and video coding and, 1027 microstrip/microstrip patch antenna and arrays, 1375-77, 1384-85, 1387-89 parabolic and reflector antenna and, 2084-86 scanning arrays, 187, 1235-36, 1239-40 scanning diversity, 731 scanning type communications, antenna arrays, 152 scatter adaptive antenna arrays and, 68-69 cellular communications channels and, 393, 394, 395, 395 free space optics and, 1851, 1855-57, 1856, 1857 indoor propagation models and, 2013, 2018-19 millimeter wave propagation and, 1271, 1434-39, 1445 optical communications systems and, 1491 optical fiber and, 1684-85, 1685, 1684, 1709, 1710, 1712, 1766, 1843, 1844, 1846 radiowave propagation and, 215

stimulated Brillouin, 1844, 1846 stimulated Raman, 1843, 1846 tropospheric scatter communication and, 2692-2704 in underwater acoustic communications, 38-40, 39, 45 wireless transceivers, multi-antenna and, 1579 scatternets, Bluetooth, 315-316 scheduling flow control, traffic management and, 1654, 1660 multimedia networks and, 1563, 1564-65 packet switched networks and, 1908-09 wavelength division multiplexing and, 656-657 wireless packet data and, 2986-87, 2986 Schmetterling antenna, in television and FM broadcasting, 2517-36 Schnorr identification protocol, authentication, 614 Schottky photodetectors, 1002, 1002 Schrodinger equation (nonlinear), in solitons, 1765-66 Schwartz-Yeh approximation, 450-451 scintillations, 1436, 1861-63, 1861 score and scores vector, maximum likelihood estimation, 1339 scrambling cdma2000 and, 362 code division multiple access and, 2876, 2877 guided, 579 Hadamard coding and, 934-935 orthogonal frequency division multiplexing and, 1876 synchronous digital hierarchy and, 2499-2500 wideband CDMA and, 2878 scrambling sequences, polyphase sequences, 1975, 1976 Scripps Institution of Oceanography, acoustic telemetry, 24 SDLC, 546 sea clutter, radar, 429-430, 429 seams in orbits, 1250 seasonal variations in radiowave propagation, 1477, 2063 seawater effects on radiowave propagation, 2064 Seaweb (see also shallow water acoustic networks), 2212-18, 2213 SECAM standard high definition TV and, 966-979 terrestrial digital TV and, 2546 second generatioin wireless systems, 370-371, 377-383, 1479, 1350, 1482-83, 2192 sectored cells, 450 sectorization, in cochannel interference, 454 sectorization of cells, in wireless multiuser communications systems, 1603 secure hash algorithm, 612 secure MIME, 1651, 2202 secure sockets layer, 1651 security, 1644-52, 1644 access control in, 1153, 1648-51 active attacks in, 1646-47 ad hoc wireless networks and, 2893-95 application level security, 1155 application programming interfaces and, 1651 asynchronous transfer mode and, 1154 audits in, 1650 authentication coding and, 218-224, 1647, 1649 authorization in, 1647 Bluetooth and, 316 bugs and software errors in, 1645 cable modems and, 335 communication security concepts in, 1651 cryptography and encryption (see cryptography) data confidentiality in, 1648 data integrity and, 1648, 1649 demilitarized zones in, 1650 denial/degradation of service in, 1646, 1647 digital signatures in, 1649 eavesdropping and, 2810 encipherment (see also cryptography), 1648-49 Ethernet, 1646 event detection in, 1650 extranet, 1165-69 firewalls in, 1650-51 general packet radio service and, 875

security (continued) global system for mobile and, 916-917 hijacking and, 1646, 2810 internet control message protocol and, 1646 Internet Security Association and Key Management Protocol and, 2813-14 Internet, 1151-57, 1165-69, 1650-52 intrusion detection and response in, 1651-52 IP over ATM and, 1154 IP telephony and, 1180-81 IPSec, 1153-54, 1651 Kerberos and, 1155 labeling in, 1649 link layer security in, 1153 MAC addresses in, 1646 man in the middle attacks and, 2810 mechanisms for, 1648-49 mobility portals and, 2194-95 multicasting, 1154-55 network layer security in, 1153-54 nonrepudiation in, 1648 notarization in, 1649 Oakley, 2813 optical fiber and, 2614 OSI reference model and, 1647-50 passive attacks in, 1645-46 pervasive mechanisms for, 1649-50 physical, 1645 privacy enhancing technologies and, 1646 promiscuous mode operation and, 1646 recovery, 1650 routing control in, 1649 satellite onboard processing and, 482 secure sockets layer and, 1651 session initiation protocol and, 2202 signature sequence for CDMA and, 2276 SKEME, 2813 software radio and, 2307, 2308, 2313 spoofing and, 1646, 2809-10 switched networks and, 1646 threats and attacks in, 1645-47 traffic analysis attack in, 1646 traffic padding in, 1646, 1649 transport layer security and transport layer security, 1651 transport layer security in, 1154-55 transport protocols for optical networks and, 2617 trusted entities in, 1649 tunneling and, 1651 virtual private networks and, 1165-69, 2809-14, 2809 wireless, 1155-56, 1646 wireless application protocol and, 2194-95 wireless communications, wireless LAN and, 1287 - 88wireless packet data and, 2985 wiretapping in, 1646 security labels, 1649 seek time, hard disk drives, 1320-21 segmentation automatic speech recognition and, 2377-78, 2377 transport protocols for optical networks and, 2617 segments, TCP, 541, 544, 2604 selectable mode vocoder, 2827-28 selected mapping, in peak to average power ratio, 1949 selection combining, wireless, 2920-21 selective acknowledgement, 1662, 2120 selective cell discarding, ATM, 206 selective filters, 1723 selective reject ARQ, 545 selective repeat, ARQ, 228, 228, 229-230 self-clocked fair queuing, in flow control, traffic management, 1660 self-electrooptic effect device, 1967, 1968 self-excited linear pulse, in speech coding/synthesis, self-healing rings, 750, 751, 1635, 1637, 1638, 1716, **1716**, 1910, 2464–68, 2495 self-interference (see intersymbol interference) self-organizing map, 1678, 2130 self-optimization, in underw/in underwater acoustic communications, 45

self-organization, neural networks, 1678 self-orthogonal coding, convolutional, 2582-84 self-phase modulation optical communications systems and, 1489, 1489 optical fiber and, 1684, 1686, 1686, 1844, 1846, 1974 solitons and, 1765 self-similar processes, traffic modeling, 1669 semianalytical MC technique, 2293-94 semiblind channel estimation, 402, 404-407 semiblind constant modulus algorithm, 2338 semicircular array antenna, 75-77, 75, 76 semiconductor lasers (see also lasers), 1777-78, 1777 semiconductor optical amplifiers, 706, 756, 1760-63, 1760, 1781, 1785, 1826, 1842, 2273, 2869 sensation level, in speech coding/synthesis, 2363-64 sensitivity free space optics and, 1858 frequency synthesizers and, 837 multibeam phased arrays and, 1518-19 receiver, 1825, 1833-34 sensor networks, 1348 sensors intelligent transportation systems and, 503 shallow water acoustic networks and, 2206 wireless networks using, 2990-96 sequential decoding of convolutional coding, 600 additive white Gaussian noise, 2143, 2144, 2155 Algorithm A in, 2140, 2145-46 backsearch limiting in, 2153-54 binary symmetric channel and, 2143, 2144, 2146 bit error rate in, 2156, 2156 buckets, stack bucket technique in, 2149 buffer overflow and system considerations in, 2159-60 channel models for, 2042-45 codewords for, 2141 coding construction for, 2160-61 coding distance profile, 2160 coding rate for, 2142, 2149 coding tree in, 2142, 2143 column distance function and, 2142, 2158 computational cutoff rate in, 2158 constraint length for, 2142 creeper algorithm and, 2154 encoding in, 2141, 2141 erasure probability in, 2159 error detection and correction in, 2159-60 evaluation function in, 2145 Fano algorithm, Fano metric in, 2140, 2146-48, 2150-54, 2151, 2152, 2153, 2154 free distance in, 2142 Gallager function in, 2157 graphical representation of, 2140-42 Hamming distance and, 2142 hard and soft decision in, 2142-45 heuristic function in, 2145-46 maximum likelihood (ML) decoder in, 2143, 2145-47 maximum likelihood sequential decoding algorithm in, 2140, 2155–56, 2182–87 memory order in, 2141 memoryless modulation in, 2144 minimum distance in, 2142 multiple stack algorith in, 2159-60 nodes in, origin and terminal, 2142 noise and, 2144 optical distance profile in, 2160-61, 2161 parallel entry systolic priority queue and, 2149 Pareto distribution, Pareto exponent in, 2157-58, 2157 performance characteristics of, 2157 quick look-in and, 2160-61 random coding techniques in, 2157 Reed-Solomon coding in, 2158 signal to noise ratio, 2144, 2156, 2161 stack algorithm in, 2140, 2155, 2159 states in, 2142 systematic nature of, 2141 time discrete channels and, 2144 trellis and, 2142, 2144, 2154-56, 2154 undetected word error and, 2159 Viterbi algorithm and, 2140, 2156, 2161-62

serial interfaces, in acoustic modems for underwater communications, 18 serial minimum shift keying, 1464-67, 1465, 1466 serial transmission, 1494–95, 1572, 1574 serially concatenated coding, 2164-79, 2164, 2179, 2180, 2180 a posteriori probability algorithm in, 2180-81 additive white Gaussian noise, 2180 applications of, 2170–72 bandwidth and, 2180 binary PAM and, 2165 binary phase shift keying and, 2165, 2180 bit error rates and, 2180, 2181–89, **2183** bootstrap effect and, 2166 code division multiple access, 2176-77, 2176 constituent coding and, 2164 continuous phase modulation and, 2173-75, 2174, 2175, 2179-90, 2180 convergence, nonconvergence region, 2166-67, 2167 design of, with interleaver, 2166-68 error detection and correction in, 2165-66, 2166, 2182-84. 2183 error floor region in, 2166-67, 2167 Euclidean distance in, 2173, 2182 example systems in, 2184-87 examples of serial, 2172–77 exit charts and, 2167, 2172 free distance in, 2167 Gray coding and, 2173, 2173 Gray mapping and, 2187, 2187, 2188 Hamming distance and, 2173, 2182 inner and outer coding in, 2164 inner CPM systems and, 2186-87, 2186 inner input weight error events and, 2187, 2187 interleaving and, 2164, 2165-66, 2165, 2183 iterative decoding in, 2180-81 log reflection coefficients and, 2175 logarithmic likelihood ratio and, 2168-72 matched coding and, 2180 maximum APP sequence detection in, 2182 maximum likelihood decoding in, 2164, 2167-68 memoryless modulation in, 2173 minimum shift keying and, 2182 multiuser interfered channel and, 2176-77, 2176 nonrecursive inner encoder in, 2170-71, 2171 normalized squared Euclidean distance and, 2182-84 outer convolutional coding in, 2185-86, 2185 parallel, 2164 phase shift keying and, 2173, 2173 power spectral density and, 2187-88 probability density evolution in, 2167 recursive systematic convolutional coding and, 2182 Reed-Solomon coding and, 2164 serial concatenation of outer encoder and inner mod-ulator in, 2172–73, **2172** serial vs. parallel, 2172, 2172 serially, 2164 serially, iterative algorithms and, 2164-78 Shannon or channel limit in, 2171 signal to noise ratio, 2165, 2166, 2181-83 simulated coding gain vs. iteration number in, 2170, 2171 sliding window SISO in, 2180-81 soft input/soft output procedure for, 2168-72, 2169, 2180-81 transfer function bound in, SCCPM and, 2182 trellis coding and, 2164, 2180 turbo coding and, 2164, 2176-77, 2176 Viterbi algorithm and, 2164 waterfall region in, 2166-67, 2167 series feed antenna arrays, 166 serrasoid technique, in pulse position modulation, 2033 servers cell planning in wireless networks and, 377, 378, 379 shallow water acoustic networks and, 2217-18, 2218, 2217 streaming video and, 2433 wavelength division multiplexing, 650-657, 650-656 service access points, cdma2000, 364 service facility, traffic engineering, 492 service ID, cable modems, 324, 335

wireless transceivers, multi-antenna and, 1579,

service level agreements, 115, 270, 668-77 service specific connection oriented protocol, 2616, 2619-20 services via mobility portals (see mobility portals and services) serving GPRS support node, 867-876, 2983-84, 2983, $\bar{2}988$ session, in flow control, traffic management, 1653 session description protocol, 2197, 2198, 2979 session initiation protocol, 2196-2206, 2979-80 applications, usage for, 2197-98 architecture for, 2199 back to back user agents in, 2198 cable modems and, 2197-98 call processing language and, 2203 caller preference setting in, 2203 cdma2000 and, 2198 conferencing in, 2202 configuration of networks for, 2204 domain name servers and, 2199 dual tone multifrequency and, 2198 elements in, 2198 emergency services and, 2203-04 ENUM mechanism in, 2198 Ethernet and, 2197 extending, 2201, 2202 forking in, 2198 gateways in, 2198 global system for mobile and, 2198 hypertext transfer protocol and, 2199, 2203 instant messaging and, 2203 IP addressing and, 2197 IP telephony and, 1174, **1175**, 1175, 1181 ISUP and, 2197, 2198 JAIN, 2203 keep alive in, 2202 layers of, 2198 locating users and servers in, 2199 MEGACO/H.248 and, 2198 mobile communications and, 2196 multimedia and, 2196-2206 multiparty calls using, 2202-03 performance, 2203 phones using, 2197-98, 2197 programming services in, 2203 proxies for, 2198 proxies in, 2197, 2201 quality of service, 2196, 2203 related protocols to, 2198 requests and responses in, 2200 routing in, 2200-01 RTSP and, 2198 secure MIME and, 2202 security in, 2202 session description protocol and, 2197, 2198 SIPstone and, 2203 soft switches and, 2198 standards for, 2197 stateful or stateless proxies in, 2198 third-generation wireless systems and, 2198 third party call control in, 2198 transactions in, 2198 transmission control protocol and, 2197 transport layer security and, 2202 universal resource identifier in, 2196, 2198 user agents in, 2196, 2198, 2201 user datagram protocol and, 2197 voice over IP and, 2197, 2198 session layer, 540, 1911 session management, in general packet radio service, 869-870 settop box, patching, 234, 234 SHA-1, 218 shadow filters, 10 shadowing, 781 cellular communications channels and, 394-395 cochannel interference and, 449 mobile radio communications and, 1481 path loss and, 1937 power control and, 1983 radio resource management and, 2093

satellite communications and, 1223, 1227-28 wireless and, 2922 shallow water acoustic networks, 2206-21 acoustic local area networks, 2212 ad hoc on demand distance vector in, 2211 ALOHA protocols and, 2208, 2209 automatic repeat request in, 2207-08, 2210-12 autonomous ocean sampling network in, 2211-12 autonomous underwater vehicles in, 2206, 2211-12 bandwidth and, 2207 carrier sense multiple access and, 2209-10, 2212 code division multiple access, 2208, 2209, 2215, 2218 congestion control in, 2211 constraints in, 2206 cyclic redundancy check in, 2207 data link control layer and, 2207 destination sequence distance vector in, 2211 digital signal processing and, 2207 direct sequence CDMA in, 2209 Doppler effect and, 2207 dual busy tone multiple access in, 2210 dynamic source routing in, 2211 EMMA probe and, 2206 energy consumption in, 2208, 2209 error detection and correction in, 2207 evolution of, 2211-12 filtering in, 2207 frequency division multiple access (frequency division multiple access and, 2208, 2215 frequency shift keying and, 2207 GEOSTAR probe and, 2206 handshake protocols in, 2215-17, 2216 intersymbol interference and, 2207 layers of network in, 2207-11 logical link control layer and, 2208 MACAW protocol in, 2210 media access control and, 2208, 2209-10, 2215-17, 2216 multiaccess interference in, 2215 multipath and, 2207 multiple access methods for, 2208-08 multiple access with collision avoidance in, 2210, 2212 near far problem and, 2208 network layer in, 2217–18, **2218** packet radio networks in, 2212 peer to peer networks in, 2208 phase locked loops in, 2207 phase shift keying and, 2207 protocols for, 2206 RAKE filters in, 2209 RF links for, 2206 routing in, 2208, 2211 Seaweb in, 2212-18, 2213 sensors and, 2206 servers for, 2217-18, 2218 spread spectrum in, 2216-17 telesonar modems in, 2215 temporally ordered routing algorithm in, 2211 time division multiple access (time division multiple access and, 2208, 2212, 2215 topologies for, 2208 virtual circuit switching in, 2208 virtual circuits in, 2211 Shalvi-Weinstein algorithm, in blind equalizers, 292 Shannon cover, constrained coding techniques for data storage, 575 Shannon or channel capacity, 2179 discrete multitone and, 745-746 low density parity check coding and, 1308 multiple input/multiple output systems and, 1453-55, 1454 rate distortion theory and, 2069 serially concatenated coding and, 2171 trellis coded modulation and, 2623 trellis coding and, 2636-37 ultrawideband radio and, 2760-61 very high speed DSL and, 2770-71 wireless and, 2915 wireless multiuser communications systems and, 1605-08, 1606

1580 - 82Shannon, Claude, 262, 275, 458, 606, 634, 1308, 2069, 2703 Shannon's law, constrained coding techniques for data storage, 573 Shannon's theory, 218 continuous phase modulation and, 592 convolutional coding and, 605 cyclic coding and, 617 Shannon-Fano coding, 634-635 Shannon-Hartley capacity theorem, 326 shape adaptive transforms, in image and video coding, 1041 - 42shape-gain vector quantization, 2127 shaping gain, in shell mapping, 2221 shared risk link group, 2463-64 SHARES high frequency communications, 948 shell mapping, 2221-27 applications and practical considerations for, 2226-27, 2226 constellation shaping and, 2221-22, 2221 cost identification in, 2224 counting sets in, 2223 decomposition in, 2224-26, 2225, 2226 generating function in, 2222, 2223 indexing algorithm in, 2223 mapping algorithm in, 2223-24 modems and, 2222, 2227 quadrature amplitude modulation and, 2221-27, 2221 shaping gain in, 2221 shift registers BCH coding, binary, and, encoding circuit for, 246-247, 246 bit interleaved coded modulation and, 278-279 convolutional coding and, 598 cyclic coding and, encoders/decoders for, 619-620 shim headers, multiprotocol label switching, 1594 shooting-bouncing ray approach, path loss, 1942 short backfire antenna, 198, 198 short leap shared protection, in optical Internet, 2466-68 short message service, 866, 2190 short range communications, intelligent transportation systems, 506 short range dependent models, traffic modeling, 1667-68 short wave (see also high frequency), 49, 1477, 2067 shortened BCH coding, binary, 246 shortest path tree, ad hoc wireless networks, 2891 shot noise, 1835, 1843 shunt feed antenna arrays, 166 side constraints, in acoustic echo cancellation, 4-5 sidebands amplitude modulation and, 133 optical fiber and, and two-tone products in, 1687 optical receivers and, 1826 Sidelnikov bound, Gold sequences, 901, 2543, 2544, 2545 sidelobe level antenna, 144, 184 sidelobes antenna and, 169 chirp modulation and, reduction in, 443-444 generalized sidelobe canceler in, 1889-90, 1889, 1892-93 minimum shift keying and, 1457 multibeam phased arrays and, 1517 packet rate adaptive mobile receivers and, 1889-90, 1892-93 parabolic and reflector antenna and, 2081 sigma delta converters, 2227-47, 2228 additive white Gaussian noise, 2237-38, 2238 amplitude input range for, 2242-43, 2242, 2243 analog to digital conversion and, 2227-47, 2228 applications for, 2243-46 architectures for, 2232 bandpass modulators using, 2238–40, **2239**, **2240** cascade converters in, 2234–35, **2234**, **2235**, **2236–37** data preprocessors using, 2243, 2244 DC canceller using, 2243-45, 2244, 2245, 2246 digital to analog conversion and, 2227-47, 2228 digital to digital conversion and, 2227-47, 2228

sigma delta converters (continued) error detection and correction in, 2229-30, 2229 feedback loops for, 2230, 2232, 2233–47 filtering in, 2228, 2232–35 finite impulse response filters and, 2228 linear model for, 2241, 2241 model for, 2230-31, 2230 multiple sample and hold converter in, 2234 noise and, 2229, 2231–32, **2231**, **2232** noise prediction loops in, 2235-38 noise transfer function and, 2231, 2231, 2232, 2233-47, 2234 signal to noise ratio, 2227, 2229, 2231 signal transfer function in, 2233 sigmoidal activation function, in neural networks, 1676 signal constellations, in orthogonal frequency division multiplexing, 1945 signal processing in acoustic modems for underwater communications, 18-19, 19 adaptive antenna arrays and, 68 magnetic storage and, 1326-33, 1327 optical fiber and, 1808 in underwater acoustic communications, 41-43, 41 signal processing for magnetic recording channels (see also magnetic recording systems), 2247-68 signal quality monitoring, 2269-74 amplified spontaneous emission and, 2272 amplitude shift keying and, 2273 analog parameters for, 2270-73 arrayed waveguide gratings in, 2271-72 bandpass filters and, 2272 bit error rate and, 2269, 2270 bit interleaved parity in, 2269 cross gain modulation in, 2273 dense WDM and, 2271-73 digital parameters for, 2269-70 digital wrappers in, 2269 eribum doped fiber amplifiers and, 2273 error detection and correction in, 2269 Ethernet and, 2269 exclusive OR gates and, 2270 fast Fourier transform and, 2272 fault isolation and, 2269 filtering in, 2272 forward error correction in, 2269 histogram evaluation in, 2270-71, 2271 lasers and, 2273 Mach-Zehnder interferometer and, 2273 optical transport networks and, 2269 phase shift keying and, 2273 photodetectors and, 2271 pilot tones in, subcarrier multiplxing, 2272-73, 2272 polarization beam splitter and, 2272 Q factors in, 2270–71, 2270 scalability and, 2269 semiconductor optical amplifiers and, 2273 signal to noise ratio and, 2269, 2271-72 simplicity of, 2269 SONET and, 2269 synchronous digital hierarchy and, 2269 transparency and, 2269 variable decision circuits in, 2269-70, 2270 wide area networks and, 2269 signal regeneration, optical, 1759-64 signal space, in pulse amplitude modulation, 2025 signal subspace, in blind multiuser detection, 301 signal to interference plus noise ratio, 1116 blind multiuser detection and, 302, 306 code division multiple access and, 458, 459-460 orthogonal frequency division multiplexing and, 1874 packet rate adaptive mobile receivers and, 1886, 1895, 1898, 1902 space-time coding and, 2324 tropospheric scatter communications and, 2697 wireless transceivers, multi-antenna and, 1579 signal to interference ratio, 1983-86 admission control and, 121 broadband wireless access and, 319 cellular telephony and, 1480-81 intelligent transportation systems and, 505, 505

local multipoint distribution service and, 319 multiple input/multiple output systems and, 1452 radio resource management and, 2090, 2091-93, 2092 Universal Mobile Telecommunications System and, 386, 387 wireless IP telephony and, 2933 signal to noise ratio in acoustic modems for underwater communications, 15, 16, 17 acoustic telemetry in, 22 adaptive antenna arrays and, 187 adaptive equalizers and, 88 admission control and, 121-122 angle modulation methods and, 820-823 antenna arrays and, 143 asymmetric DSL and multimedia transmission in, 1573 bit interleaved coded modulation and, 277, 278, 285 blind equalizers and, 297 broadband wireless access and, 321 cable modems and, 326, 327-328, 327 in channel modeling, estimation, tracking, 413, 414 chaotic systems and, 422, 425, 425 chirp modulation and, 442, 445-447 community antenna TV and, 514-517, 519, 520-522, 526 continuous phase frequency shift keying and, 596-597, 597 continuous phase modulation and, 588-589, 2181-83, 2189 convolutional coding and, 602-605 discrete multitone and, 745 diversity and, 731 expectation maximization algorithm and, 773-774, 774, 775 fading and, 786-788 free space optics and, 1858-62, 1859 image compression and, 1062 linear predictive coding and, 1264 location in wireless systems and, 2964-65 magnetic recording systems and, 2264 magnetic storage and, 1326, 1327, 1331 matched filters and, 1337 microwave and, 2567, 2571 multibeam phased arrays and, 1519 multiple input/multiple output systems and, 1453-54 optical communications systems and, 1485-87, 1493 optical fiber systems and, 1709, 1825, 1841, 1846-48 optical transceivers and, 1837 orthogonal frequency division multiplexing and, 1873 packet rate adaptive mobile receivers and, 1898 phase shift keying and, 714 photodetectors and, 997 photonic analog to digital conversion and, 1965 polyphase sequences and, 1975 powerline communications and, 2004 predistortion/compensation in RF power amplifiers and, 530 pulse amplitude modulation and, 2026 pulse position modulation and, 2039 quadrature amplitude modulation and, 2053-55 rate distortion theory and, 2069, 2070 satellite communications and, 1224, 1230 sequential decoding of convolutional coding and, 2144, 2156, 2161 serially concatenated coding and, 2165, 2166 serially concatenated coding for CPM and, 2181-83, 2189 sigma delta converters and, 2227, 2229, 2231, 2235 signal quality monitoring and, 2269, 2271-72 software radio and, 2314 space-time coding and, 2325, 2327, 2329 spatiotemporal signal processing and, 2335-36, 2338-39, 2339 speech coding/synthesis and, 1305, 2342, 2345-46, 2346, 2353-55, 2361, 2368 spread spectrum and, 2395 synchronization and, 2473-85 tapped delay line equalizers and, 1690 terrestrial digital TV and, 2547

trellis coded modulation and, 2623-24, 2628, 2630, 2634 trellis coding and, 2637-39, 2642 tropospheric scatter communications and, 2697 turbo equalization and, 2716-27 turbo trellis coded modulation and, 2738 ultrawideband radio and, 2756 in underwater acoustic communications, 37-38 waveguide and, 1416 wavelength division multiplexing and, 2869 wireless and, 2915, 2919, 2921 wireless infrared communications and, 2927 wireless multiuser communications systems and, 1606. 1619 wireless transceivers, multi-antenna and, 1579, 1580, 1582, 1583, 1584 signal to quantization noise, 1966, 2830, 2833, 2837 signal transfer function, sigma delta converters, 2233 signaling ATM and, 204, 204, 2090-14 cdma2000 and, 359 global system for mobile and, 913-916, 913 wavelength division multiplexing and, 653-654 signaling ATM adaptation layer protocol, 204 signaling layer, multiprotocol label switching, 1595-97 signaling radio burst protocol, 359, 364 signaling system 7, 113, 906 signature sequences, CDMA, 2274-85, 2275, 2274 autocorrelation and, 2276-85 average interference power and, 2283 common types of, 2278-83 cross correlation in, 2276-85 direct sequence CDMA and, 2274, 2283, 2284 filtering in, 2275 frequency hopping CDMA and, 2276 Gold sequences as, 2281-82, 2281, 2282 intersymbol interference and, 2278, 2283 Kasami sequences in, 2282 matched filters in, 2275 maximal length sequences in, 2279-81, 2280 multiple access interference and, 2278, 2283 periodic sequences in, 2279 polyphase sequences and, 1975 random sequences and, 2276 receivers and transmitters for, 2275-76, 2275 security and, 2276 spread spectrum and, 2276-78 Walsh-Hadamard sequences in, 2282-83 wideband and, 2282, 2283 silica glass, in optical fiber, 434 Simmons' bounds, in authentication coding, 219-220 simple mail transport protocol, 541, 544 simple merging piggybacking, 233 simple network management protocol, 200, 544 simplex coding, 932 simplex method, in antenna arrays, optimization using, 161 simplicity, flow control, traffic management, and Occam's razor concept, 1653 simulated annealing, 161-161, 2130 simulation cellular communications channels and, 396-397 high frequency communications and, 955-956 very high speed DSL and, 2788-89 simulation of communication systems, 2285-95, 2285 bandpass inputs and, 2287 bandpass random processes, sampling of, 2286–87 bandpass signals, sampling of, 2286 bit error rate and, 2293-94 digital signal processing and, 2285 discrete channel model for, 2291 discrete Fourier transform and, 2287, 2288 discrete time representation of, 2285-87 domain convolution in, 2287 fading, 2290-91, 2291 fast Fourier transform, 2288 finite impulse response filters and, 2287-88, 2292-93 first zone output components and, 2289 frequency division multiplexing and, 2286 functional blocks and, 2287-91

hardware description language and, 2285

simulation of communication systems (continued) infinite impulse response and, 2288, 2288 linear time invariant components and, 2287-88 linear time varying components and, 2288-89 lowpass random processes, sampling of, 2286-87 lowpass signals, discrete time representation of, 2286 Markov model for, 2291, 2291 memory nonlinearities and, 2290 memoryless nonlinearities in, 2289-90 mobile communications channels and, 2290-91, 2290 Monte Carlo, 2285, 2291-94 multipath, 2290-91 nonlinear components and, 2289-90 performance measurement using Monte Carlo, 2293-94 power spectral density and, 2287, 2291, 2292-93 random number generation and, 2291-93 random processes and, uncorrelated and stationary, 2290 sampling in, 2287, 2294, 2294 semianalytical MC technique in, 2293-94 swept power measurements in, 2290 tapped delay lines and, 2289, 2289 Viterbi algorithm in, 2287 waveform level, 2285, 2285 SincGars, 2310 sine wave speech, 2362 single carrier frequency domain equalization, 2329-30, 2330 single channel per carrier frequency division multiple access (frequency division multiple access and, 825 satellite communications and, 878 trellis coding and, 2638 single frequency networks digital audio broadcasting and, 679, **679** terrestrial digital TV and, 2554–55 single hop WDM networks, 1551 single input multiple output systems blind equalizers and, 292, 294–296 in channel modeling, estimation, tracking, 400, 400, 403-407 multiple input/multiple output systems and, 1451 wireless transceivers, multi-antenna and, 1582-83 single input/single output systems blind equalizers and, 287-288, 288, 291, 293 multiple input/multiple output systems and, 1451 space-time coding and, 2327 single letter fidelity, rate distortion theory, 2072-76 single link admission control, 117 single mode optical fiber, 1507, 1707, 1842, 1845, 1848 single parity check coding, 1540-43, 2007 single phase unidirectional transducer, 2450, 2450 single sideband, 36, 1478, 1826, 1930 single sideband AM, 133, 135-136, 136, 140, 140 singletalk, acoustic echo cancellation, 4 singular value decomposition, blind multiuser detection, 301 sinusoidal coding, in speech synthesis/coding, 1300-01 sinusoidal frequency shift keying, 1457, 1458, 1459, 1462, 1473–74 SIPstone, 2203 64B/66B encoding, Ethernet, 1508 SKEME, 2813 skew, multimedia networks, 1562 skip distance, 2065, 2067 skip fading, 2065 sky wave propagation, 208, 946-958, 2060-65 skyphone, 2824 skyscraper broadcasting, 236 slant range, in satellite communications, 1225-26, 1226 slave (see master slave configuration) sliding block decoders, 575, 2254 sliding window flow control, 545 sliding window, 227, 2604, 2609-11, 2609 sliding window RLS, 415 sliding window SISO, 2170, 2180–81, 2180 slot antenna, 142, 169, 170, **170**, 180 slot time, Ethernet, 1281 slots, 340 slotted antenna arrays, 142

1559 slotted waveguide, millimeter wave antenna, 1428, 1429 slotted waveguide array, 1391, 1391, 1392, 1391 slow fading path loss and, 1937 simulation and, 2290-91 tropospheric scatter communications and, 2698 wireless multiuser communications systems and, 1604 smart antenna systems, 163, 180, 184, 187, 191, 1580 smart messaging, 2899 smartphones, 2191 smoothing, in cryptography, 610 smoothing factors, in blind multiuser detection, 303, 304 snapshot policy piggybacking, 233 Snell's laws of reflection, 210, 2082, 2086, 2559 sockets, 541 sofic systems, in constrained coding techniques for data storage, 573-575 soft and safe admission control, 2094 soft bit error rate, 474 soft decision decoding BCH coding, binary, and, 251-252 BCH (nonbinary) and Reed-Solomon coding, 261 convolutional coding and, 601-602, 602 low density parity check coding and, 1309, 1312 magnetic recording systems and, 2257 multidimensional coding and, 1541 Reed-Solomon coding for magnetic recording channels and, 475 sequential decoding of convolutional coding and, 2142-45 trellis coding and, 2640 soft handoffs, 366, 2093 soft in/soft out decoders, 560, 564-567 soft input soft output systems serially concatenated coding and, 2168-72, 2169, 2180-81 soft output decoding algorithms and, 2295, 2297, 2302 turbo coding and, 2713, 2714, 2728-37 turbo equalization and, 2716-27 soft output decoding algorithms, 2295-2304 additive white Gaussian noise, 2295, 2296 BCJR algorithm and, 2295, 2297 BCJR algorithm and, 2299-2301, 2299 binary phase shift keying and, 2296 forward error correction and, 2295-96 iterative coding and, 2295 iterative decoding and iterative decoding, 2302 logarithmic maximum a posteriori algorithm in, 2301-02 maximum a posteriori algorithm and, 2295, 2297 maximum logarithmic MAP algorithm in, 2302 soft input soft output systems and, 2295, 2297, 2302 soft output Viterbi algorithm and, 2295, 2297-99, 2298, 2302 system model for, 2295-2297, 2296 trellis coding and, 2296, 2296 turbo coding and, 2295, 2302, 2303 Viterbi algorithm and, 2295, 2297-99, 2298, 2302 soft output Viterbi algorithm, 2295, 2297–99, **2298**, 2302 soft switches, 2198 software communications architecture, 2311-12, 2311 software defined radio, 2304, 2305, 2309-12 software radio, 2304-24 alternative technologies for, 2315-16 analog to digital conversion in, 2305, 2306, 2308, 2313 application programming interfaces (API) for, 2311 applications for, 2307, 2318–21 automatic link establishment protocol in, 2313–14 bandwidth and, 2315-17 beamforming and, 2307 benefits of, 2304 bit error rate and, 2314 CASE tools and, 2305 certification in, 2319-21 code division multiple access, 2312, 2312, 2314, 2316

slotted ALOHA, 128, 341-342, 500-501, 500, 1552,

cognitive radio and, 2307 common object request broker architecture and, 2304, 2310, 2311 core framework, 2312 development parameters and risks in, 2317-18 digital modular radio in, 2306 digital signal processing and, 2304, 2306, 2316 digital to analog conversion and, 2305, 2306, 2308, 2313 environment management streams in, 2313-14 extensible markup language and, 2304 field programmable gate arrays and, 2307, 2316 frequency division multiple access and, 2312, 2312 functional model of, 2306–08, 2307 functions, components, design rules for, 2310 GFLOPS processing in, 2311 global system for mobile and, 2314 history and development of, 2305 horizontal architecture design rules in, 2310 host platforms for, 2304 ideal design of, 2305-09, 2304 industry standard architectures for, 2311-12 InfoSec and, 2307, 2308, 2313 interfaces for, 2305, 2308-09, 2308 interference and, 2306 joint tactical radio system in, 2306, 2311 local area networks and, 2307 mathematical structure of, 2311 modems for, 2308 multimode, 2307 multiple instruction/multiple datastream in, 2313 offline adapation and factories in, 2314 online adaptation in, mode selection/download management, 2314 orthogonal frequency division multiplexing and, 2314 personal communications systems and, 2306 personal digital assistants and, 2314 plug and play, 2310 programming languages and, 2304 public switched telephone network and, 2305 quality of service, 2307 RAKE receivers and, 2307 real time channel processing streams in, 2312, 2312 reference platform for, 2317, **2318** security and, 2307, 2308, 2313 service and network support for, 2307 signal flows in, isochronous and interdependent, 2313, 2313 signal to noise ratio, 2314 software communications architecture in, 2311-12, 2311 software defined radio and, 2304, 2305, 2309-12 software tools for, 2314-15, 2315 SpeakEasy, First, and Trust systems in, 2305, 2316 specification and description language and, 2304-05 spectrum management and, 2319-21, 2320 stability and, 2319 standards for, 2304-05 synchronous digital hierarchy and, 2307 technology of, 2304 time division multiple access (time division multiple access and, 2312-14, 2312 type certification and, 2319 unified modeling language and, 2304, 2312 upload/download process in, 2307 uploads/downloads in, 2314 vertical architecture design rules in, 2310-11 wireless LAN and, 2314 solar cycles, solar flares, in radiowave propagation, 2060-61, 2066, 2067 solid immersion lens technologies, optical memories, 1739 solid state memories, 1319 solid state power amplifiers, 531, 535 solitary waves, 1764 solitons (see also optical fiber), 1764-73 acoustic jitter in, 1767 amplification and loss and, 1767 breathers in, 1766 chirped fiber gratings and, 1768 control of, 1768

solitons (see also optical fiber) (continued) conventional types, 1765 dense WDM and, 1771 discovery of solitary waves and, 1764 dispersion and, 1764, 1765 dispersion compensating fibers in, 1768 dispersion managed types, 1765 dispersion managed, 1768-70, 1769 electrabsorption modulators in, 1770, 1771 erbium doped fiber amplifiers and, 1764 evaluation and future of, 1771 experiments and field trials of, 1770-71, 1771 fiber ring lasers and, 1771 gain switched lasers and, 1771 Gordon-Haus jitter in, 1767, 1769 group velocity dispersion in, 1764, 1765, 1769 intrachannel cross phase modulation in, 1769-70 intrachannel four wave mixing in, 1769-70 intrachannel impairments and, 1769-70 inverse scattering transform in, 1766 laser sources for, 1770 lasers and, 1764 loop experiments in, 1770 optical fiber and, 1686, 1714, 1848 optical sources for, 1770 polarization and, 1768 polarization mode dispersion and, 1768, 1770 polarization multiplexing in, 1771 refractive index and, 1764 scattering and, 1766 Schrodinger equation (nonlinear) in, 1765-66 self-phase modulation in, 1765 third order dispersion and, 1766-67 timing jitter and, 1767 wave propagation in optical fiber and, 1765 wavelength division multiplexing and, 1768, 1769 sonar, 29, 32, 35, 35 SONET, 1798, 2461, 2485-2510, 2486, 2485, 2509 add drop multiplexers in, 2493-94, 2494 ATM and, 201, 273 automatic protection switching in, 2494-95, 2495 bidirectional path switched ring in, 2495-96 concatenation of STS in, 2488 dense WDM and, 748-757, 749 Ethernet vs., 1501, 1512 fault tolerance and, 1634, 1635 free space optics and, 1851 Gigabit Ethernet and, 1509 history and development of, 2485-87 matched nodes in, 1637, 1637 optical cross connects/switches and, 1701, 1782, 1798 optical Internet and, 2464-68 overhead levels in, 2488-93 packet switched networks and, 1910 payload mapping in, 2493 payload pointer in, 2489-92, 2492 reliability and, 1634, 1635 ring topologies in, 2495-96, 2496 self-healing rings in, 2495 signal quality monitoring and, 2269 signal structure in, 2487-89, 2487, 2488 SPE assembly/disassembly in, 2493 unidirectional path protection/path switched rings in, 2495 virtual tributaries in, 2488-89, 2491 wavelength division multiplexing and, 2838-46, **2838**, **2839**, 2863–73 Song-Park polyphase sequences, 1979 sound generation (see propagation of sound) sound pressure, 31 sound pressure level, 32, 2366-67, 2366 source address table, Ethernet, 1505-06 source allocation protocol, 1552 source bits, constrained coding techniques for data storage, 573, 575 source coding compression and, 631 information theory and, 1111 rate distortion theory and, 2069 transform coding and, 2593-94, 2594 source models, in traffic modeling, 1671-73, 1671

source routing, in multimedia networks, 1566 space and time optimization, antenna arrays, 163-164, 163, 164 space communications protocol standards, 1233 space diversity, 1481, 2564 space diversity antenna, 190 space division multiple access antenna for mobile communications and, 191 spatiotemporal signal processing and, 2336 waveguide and, 1416 wavelength division multiplexing and, 2863 space division switching, ATM, 202-203 space factor, in waveguide, 1419 space wave propagation, 208 space waves, leaky wave antenna, 1237 space-time coding, 2324-32, 2324, 2324 additive white Gaussian noise, 2325, 2326 antenna and, 2324 bandwidth in, 2327 BCJR decoders and, 2328 bit error rate and, 2330 channel estimation issues in, 2330 channel impulse response and, 2327 channel model for, 2327-28 coherence bandwidth in, 2327 data rates and, 2324 decision feedback equalizers and, 2328 diversity and, 2324 diversity gain and, 2324 Doppler frequency and, 2325 equalization of, on frequency selective channels, 2327 - 30expectation maximization algorithm and, 773 fading and, 2324 global system for mobile and, 2326 interference and, 2324 intersymbol interference and, 2327 maximum a posteriori decoders and, 2328 maximum likelihood detectors and, 2325, 2327 maximum likelihood sequence estimation and, 2328, 2329 multipath and, 2324 multiple input/multiple output systems and, 2327, 2330 narrowband communications and, 2324 orthogonal frequency division multiplexing and, 2328-29, 2329 peak average ratio in, 2330 perfect root of unity sequences in, 2330 phase shift keying and, 2326, 2326 power spectral density and, 2325 signal to interference plus noise ratio and, 2324 signal to noise ratio, 2325, 2327, 2329 single carrier frequency domain equalization in, 2329–30, **2330** single input/single output systems and, 2327 space time block coding in, 2326–27, 2326, 2329–30 space-time trellis coding in, 2325-26 time reversal space time block coding in, 2329, 2330 trellis coding and, 2328-29 turbo coding and, 2328-29 Viterbi algorithm and, 2326 Spaceway, 2112 spacing, in equalizers, symbol-spaced vs. fractionally spaced, 86-87, 87 SPADE system, 825 spanning trees, multicasting, 1532 spark gap transmitters, 48-49, 48 sparse light splitting, routing and wavelength assignment in WDM, 2105 sparse multipath channels, tapped delay line equalizers, 1688-96 spatial diversity, 1230-31, 1603 spatial division multiple access, 163, 455 spatial multiplexing, MIMO systems, 1452, 1455 spatial processing, antenna arrays, 163 spatial renewal process model, traffic modeling, 1666, 1668 spatial reuse, media access control, 1344 spatiotemporal signal processing, 2333–40, 2333 antenna and, 2333–40, 2333

beamforming antenna and, 2333-40, 2333 BLAST architecture and, 2333 blind techniques for, 2333 block coding and, 2333 burst structure and, 2337-38, 2337 cochannel interference and, 2333 direct sequence CDMA and, 2336 discrete channel model for, 2333–34 diversity and, 2333, 2334–36 fading and, 2333-40, 2333 flat fading and, 2334-36 frequency division multiple access (frequency division multiple access and, 2336 interference and, 2337-38 intersymbol interference and, 2333, 2336 maximal ratio combining and, 2336 mean square error and, 2338-39, 2339 multipath and, 2333 multiple access interference and, 2333, 2336 pseudotraining symbols, 2338 semiblind channel equalization in, 2336-38 semiblind constant modulus algorithm and, 2338 signal to noise ratio, 2335-36, 2338-39, 2339 simulation examples and, 2338-39 space division multiple access and, 2336 time division multiple access (time division multiple access and, 2336 training and, 2333, 2338 trellis spacetime coding and, 2333 SpeakEasy, 2305, 2316 speaker verification systems, 2379-80 specification and description language, 2304-05 spectral attenuation, in optical fiber, 435 spectral broadening, fading, 784 spectral waveform coding, 2835-37 spectral costs, in cell planning in wireless networks, 374 spectral efficiency optical communications systems and, 1488 optical fiber systems and, 1848-49 wireless transceivers, multi-antenna and, 1579, 1581-82 spectral lines (Lebesgue decomposition), PPM, 2035, 2037, 2038 spectral noise shaping, photonic analog to digital con-version, 1960 spectral null constraints, optical recording, 579, 580 spectral shaping power spectra of digitally modulated signals and, 1990–91 pulse amplitude modulation and, 2024 pulse position modulation and, 2032-33 spectral thinning, concatenated convolutional coding, 558 spectrally raised cosine modulation, 585 spectrum (see frequency) spectrum efficiency, cochannel interference, 452-454 spectrum of radio waves, 208 specular effects, in radiowave propagation, 211-213, 212, 213 speech coding/synthesis, 2340-59 a priori probabilty and, 2367 acoustic echo cancellation and, 4, 5 adaptive differential PCM in, 2343, 2354, 2355, 2372, 2382, 2820–22, **2822** adaptive multirate coding in, 2355, 2828 adaptive postfiltering in, 2346 algebraic CELP in, 1304, 1306, 2349, 2355, 2356, 2826-27 algebraic vector quantized CELP in, 1306 algorithmic (objective) quality measures in, 2353-54 analog to digital conversion in, 2370 analysis by synthesis in, 2344-50, 2344, 2823-24 analysis of speech in, 2370-71 applications for, 2340 articulation in, 2364-65 articulation in, place of, 2360 articulation index in, 2362, 2363-68 artificial neural networks in, 2378-79 audibility in, 2363-64 automatic speech recognition and, 2373-79, 2382 average magnitude difference function in, 2350

videoconferencing and, 2827

vocoders (LPC type), 2350-51, 2350, 2819-29

speech coding/synthesis (continued) bandwidth in, 2363-67, 2365 bit rates in, 2341 boundary conditions, open- and closed-glottis, 2350 broadband and, 2362 cepstrum in, 2373, 2386 channel coding and, 1299 characteristic waveforms in, 1301–02, **1302** clippers for, 2362 coarticulation in, 2361 code division multiple access, 2354 code excited linear pulse in, 1266-67, 1302-05, **1303**, 2348–49, **2349**, 2372, 2382, 2820–28 comb filtering in, 2378 companded PCM, 2342, 2342 companders and, 529 comparitive measures of quality in, 2352-53 compression and, 648 conjugate structure CELP in, 1304, 1306 continuous speech recognition in, 2377 continuously varying slope delta modulation in, 2343, 2356 convolutional coding and, 2355 current technologies in, 2382 cyclic redundancy check in, 2355 diagnostic acceptability measure in, 2352 diagnostic alliteration test in, 2352 diagnostic rhyme test in, 2352 differential PCM in, 2342-43, 2343 digital to analog conversion in, 2370 dynamic time warping in, 2373 EFR algorithm in, 2827 embedded and multimode coding in, 2354-55 enhanced full rate coders in, 1306 enhanced variable rate coder in, 1306, 2827 error detection and correction in, 2342, 2355, 2367, 2372 Euclidean distance and, 2355 excitation functions in, 2341, 2347-48 feedforward/feedback algorithms in, 2354 filtering in, 2344-45, 2370, 2378 finite impulse response filter in, 2343, 2346 focused search technique in, 1306 formants in, 2361, 2820 frame-based analysis in, 2346-48, 2346 fricatives in, 2360 full rate and half rate standards for (Japanese), 909-911, 2827 fundamental frequency estimation in, 2372-73, 2820 G.723.1 multimode coder in, 2354-55 gain in, optimum, 2347-48 global system for mobile and, 909-911, 2356, 2819-20, 2827 Hamming distance and, 2355 hidden Markov models and, 2373-80, 2385 importance density function in, 2365-66, 2366 infinite impulse response filter in, 2343, 2346 intelligibility in, 2352, 2362, 2363, 2367 interactive voice response systems and, 2384 IP telephony and, 1178-79 joint position and amplitude search in, 1306 joint source channel coding and, 2355 language models for, 2376–77, 2385, 2388–89 Levinson-Durbin algorithm in, 2349 line spectral frequencies in, 2350, 2372 line spectral pairs in, 2350, 2821 linear prediction in, 2820-23, 2821, 2822 linear predictive coding and, 1264–67, **1264**, **1265**, **1266**, **1267**, 2341, 2344–50, **2344**, 2372, 2373 linear time invariant coders in, 2341 long term prediction in, 2823-24, 2825 lossy coders in, 2341 low bit rate speech coding in, 1299-08, 1299 low delay CELP in, 2349, 2355, 2825-26, 2826 masking spectrum in, 2345, 2364 mean opinion score in, 1179, 1305, 2352, 2353, 2819-20 mean square error, 2347-48 mel scale frequency cepstral coefficients in, 2373, 2382 mixed excitation linear pulse in, 1266-67, 1300, 1306, 2351, **2351**, 2356, 2822–24, **2823**

MPEG compression and, 2356, 2819 multiband excitation coding in, 2351 multipulse LPC in, 2348, 2355 narrowband coding in, 2341 network issues and, 2354-55 noise in, 2353-54, 2378 numerical measure of perceptual quality in, 2352 Nyquist rate in, 2371 Nyquist theorem and, 2370 parameter estimation from speech segments in, 1302-05, 2371 perception of speech in, 2359-69 perceptual analysis measurement system, 1179 perceptual error weighting in, 2345 perceptual speech quality measurement, 1179, 1305, 2354 phonemes in, 2360-63, 2370, 2371 pitch detectors in, 2372-73 pitch period in, 2361 pitch prediction filtering in, 2344-45, 2345 pitch synchronous innovation CELP in, 1304 predictive coding for, 1300 processing in, 2369-83 prototype waveform interpolative coding in, 2351 psychophysical (subjective) quality measurement in, 2352–53 public switched telephone network standards for, 2355 pulse coding modulation and, 1299, 2341-42, 2371, 2372 QCELP and, 2354, 2826 quality measurement in, 2351-54, 2372 quantization in, 2340-41 rapidly evolving waveform in, 1301 rate compatible channel coding and, 2355 rate compatible punctured convolution coding in, 2355 reduced acoustic information and, 2361-62 regular pulse excitation algorithm in, 2824 regular pulse excitation with long term predictor in, 1304, 2356 relative spectral method in, 2378 relaxed CELP and, 2827 residual quantization in, 2345-46 rollover effect in, 2366 sampling in, 2340-41, 2370 segmentation in, 2377-78 selectable mode and, 2827-28 self-excited linear pulse in, 2349 sensation level in, 2363-64 signal characteristic of speech in, 2360-61, 2360 signal to noise ratio in, 1305, 2342, 2345-46, 2346, 2353-55, 2361 sine wave speech in, 2362 sinusoidal coders for, 1300-01 skyphone and, 2824 slowly evolving waveform in, 1301-02, 1302 sound pressure levels in, 2366-67, 2366 speaker verification systems and, 2379-80 spectral representation of signal for, 2361, 2371-72, 2820 spectrogram of, 2361, 2383, 2384 speech intelligibility index in, 2362, 2363, 2366-67, 2366 speech recognition software and, 2370 speech transmission index in, 2362, 2367-68 standards for, 2355-56, 2819 stochastic codebooks in, 2345 stops in, 2360 subband coding in, 2343-44, 2343 temporal representation of signal for, 2360-61, 2360 text to speech systems in, 1304-05 text to speech synthesis and, 2380-82 third-generation CELP and, 2827-28 thresholds in, 2364, 2364 top down and bottom up processing in, 2363 unequal error protection in, 2355, 2764 variable rate CELP and, for CDMA, 2826 vector quantization in, 2350, 2372 vector sum excited linear pulse in, 2349, 2356, 2821, 2825, 2825

modulation and, 2368

voice activity detection and, 2355 voice over IP and, 2354, 2355, 2827 voice, unvoiced, and mixed signals in, 2820 voicing in, 2360 waveform coders for, 1301-02, 2341-43, 2819 wideband coding in, 2341 zero crossing rate in, 2371 zero state response/zero input response in, 2347 speech intelligibility index (SII), 2362, 2363, 2366-67, 2366 speech perception (see also speech coding/synthesis), 2359-69 speech recognition (see also automatic speech recognition), 2370, 2373-79 speech signals, in underwater acoustic communications, 37 speech transmission index, 2362, 2367-68 speed of sound, 15, 30 spherical conformal antenna arrays, 152-153 spherical coordinates, in antenna arrays, 142 spiders, microstrip/microstrip patch antenna, 1388, 1388 spiral antenna, 935–946, **936**–945 splices, optical fiber, 440, 1707 split state diagrams, convolutional coding, 604-605, 604 split TCP, satellite communications, 2120 splitters community antenna TV and, 519, 520 very high speed DSL and, for POTS and, 2785-86 splitting algorithms, media access control, 1347 splitting method in quantization, 2129 splitting ratio, optical couplers, 1699 spontaneous emission, lasers, 1776 spoofing, 1646, 2120, 2809–10 sporadic E radiowave propagation, 2065 spot beam satellite communications, 877-878, 1249 spotsize, of lasers, 1777 spread F, 2065 spread spectrum, 2391-2402, 2391 adaptive receivers for spread-spectrum systems, 95-112 ALOHA protocol and, 131, 132 Bluetooth and, 309-310 cdma2000 and, 359, 367 chann/in channel modeling, estimation, tracking, 409 chirp modulation in, 440-448 code division multiple access and, 458, 2276-78, 2400 commercial systems of, 2399-2400 community antenna TV and, 2399–240 direct sequence, 2392–96, **2392** diversity and, 733-734 ensemble-averaged autocorrelation in, 428 error detection and correction in, 2394-95 feedback shift registers and, 789 frequency division multiple access (frequency division multiple access and, 828 frequency hopped, 2396-99 frequency shift keying and, 2396-97 global positioning system and, 2399 global system for mobile and, 2400 Gold sequences and, 428 intelligent transportation systems and, 505 interference and, 1130-41, 2393-94 IS95 cellular telephone standard and, 347, 348, 2400 media access control and, 1343-45 packet rate adaptive mobile receivers and, 1886 performance in, 2396 phase shift keying and, 2397 polyphase sequences and, 1975 pulsed interference in, 2396 satellite onboard processing and, 482 shallow water acoustic networks and, 2216-17 signal to noise ratio, 2395 signature sequence for CDMA and, 2276-78 synchronization and, 2479-81, 2479 underw/in underwater acoustic communications, 41 wideband CDMA, 2400 WiFi and, 2400 wireless communications, wireless LAN and, 1285, 2399

spreading, wireless multiuser communications systems, 1611-12, 1612 spreading coding adaptive receivers for spread-spectrum system and, 106 cdma2000 and, 367 media access control and, 1344-45 spreading factor in channel modeling, estimation, tracking, 411 code division multiple access and, 459 signature sequence for CDMA and, 2275 spreading rates, cdma2000, 359 spreading sequence, multicarrier CDMA, 1523 spreading sequences cell planning in wireless networks and, 386 chaotic systems and, 422, 428, 428 polyphase sequences and, 1975 Universal Mobile Telecommunications System and, 386 spur free dynamic range, 1965 spur suppression, in frequency synthesizers, 858-862 square root bound authentication coding, 220 square root raised cosine filter, in cable modems, 328-329 squared error distortion, compression, 640 squaring loop, in pulse amplitude modulation, 2028, 2028 stability, 1653, 2240-43, 2319 stack algorithm, in sequential decoding, 2140, 2148-50, 2155, 2159 stacks of protocols, 541 standards acoustic echo cancellation and, 4 cdma2000 and, 358, 359-367 CDROM and, 1736, 1737 cellular communications channels and, 397 cellular telephony and, 1479 compact disc, 1736 digital versatile disc and, 1737-38 general packet radio service and, 866 Gigabit Ethernet and, 1509 global system for mobile and, 905 image and video coding and, 1048-55 image compression and, 1070 IMT2000 and, 1095-97 IP telephony and, 1181 local multipoint distribution services and, 1268-70 modems and, 1497-99 optical memories and, 1736, 1737 personal area networks and, 2683 powerline communications and, 1996–97, 2002 pulse position modulation and, 2041 satellite communications and, 2113 session initiation protocol and, 2197 software radio and, 2304-05 speech coding/synthesis and, 2355-56, 2819 terrestrial digital TV and, 2546, 2550-52 very high speed DSL and, 2779-84, 2791 vocoders and, 2819 wireless infrared communications and, 2929–30 wireless LANs and, 2682, 2945–46 wireless MPEG 4 videocommunications and, 2978 wireless, 371 standing wave linear dipoles, 1257 standing waves, in waveguide, 1394 Stanford Research Institute, 268 star topologies, optical fiber, 1716-17, 1717 STARNET, 1720 start frame delimiter, Ethernet, 1503 start frame delimiters, 1282 start/stop transmission, modems, 1495 state combination, constrained coding techniques for data storage, 578 state diagrams, feedback shift registers, 790-791, 790, 793, 794 state management, for optical networks, 2618 state matrices, in cryptography, 608 state merging, in constrained coding techniques for data storage, 575

state space, in traffic engineering, 492–493, 493 state splitting, in constrained coding techniques for data storage, weighted vs. consistent, 578–579, 578 state transition diagram constrained coding techniques for data storage and, 573 magnetic recording systems and, 2253-57, 2256 traffic engineering and, 492-495, 493, 495, 498 state variables, in low density parity check coding, 1316 states, in sequential decoding of convolutional coding, 2142 static mode analysis, incell planning in wireless networks, 388, 389 static routing and wavelength assignment, 2100-01 static SLSP, in optical Internet, 2469-70 static vs. dynamic optical WDM networks, 1798 station keeping, satellite communications, 1248 stationarity, in traffic engineering, 487 statistical multiplexing, 2420-32 adaptive receivers for spread-spectrum system and, 95 admission control in, 2428-29 bandwidth in, 2428-29 flow control and, 1625 fluid buffer models and, 2427-28 packet switched networks and, 1907-09, 1908 video and, 2424-32, 2425 steering, in active antenna, 66, 69, 70 Steiner tree problem, in multicasting, 1532-33 Steiner triple system, in low density parity check coding, 659 step size factor, in acoustic echo cancellation, 7 stereo broadcasting, FM, 823, 823, 824 stereophonic systems, acoustic echo cancellation, 11, 11 stiction, 1351 STIMAX free space grating, 1754-55 stimulated Brillouin scattering, 1844, 1846 stimulated emission, lasers, 1776 stimulated Raman scattering, 1685, 1712, 1843, 1846 stimulated scattering, 1684-85, 1685, 171 stochastic channel modeling and estimation, 405, 411-412 stochastic codebooks, in speech coding/synthesis, 2345 stochastic differential equation, in chaotic systems, 424, 425 stochastic maximum likelihood algorithm, in channel modeling, estimation, tracking, 405-406 stochastic relaxation quantization, 2130 Stoke's theorem, in active antenna, 55 Stokes photons, in optical fiber, 1712 Stokke method and groundwave propagation, 2060 stop and go algorithm, 92, 1660 stop and wait, automatic repeat request, 226, 226, 229-230, 545 stop and wait flow control, 545 stopping condition in quantization, 2129 stops, in speech coding/synthesis, 2360 storage area networks, 1714, 1733 storage devices, 1319 store and forward networks, 1907 Strat-Chu equation, 172 stream ciphers, 607-609, 608 streaming sources, in traffic modeling, 1671 streaming video, 2432-41, 2433 additive increase/multiplicative decrease in, 2439 application layer and, 2436 buffers in, 2438 client and server for, 2433, 2434 coding for, 2433-34 congestion control in, 2438-39 future of, 2439-40 MPEG compression and, 2435-36 packetization and transport layer issues in, 2436-37 real time streaming protocol in, 2438 real time transport protocol and, 2436-37 RTP control protocol in, 2438 session control in, 2437-39 transmission control protocol in, 2439 user datagram protocol in, 2439 wireless $\widetilde{\text{MPEG}}$ 4 videocommunications and, 2974 stripe, magnetoresistive, 1323 subband coding compression and, 648 speech coding/synthesis and, 2343-44, 2343

waveform coding and, 2836, 2836 subband systems, acoustic echo cancellation, 11-12, 11 subcarrier multiplexing, in signal quality monitoring, pilot tones, 2272–73, **2272** subcarriers/subchannels, in orthogonal frequency division multiplexing, 1867 subdomain basis functions, in antenna modeling, 174-176 submarines, acoustic telemetry, 24 subnetwork connection protection, 1635 subreflectors, 1920, 2083-84, 2083 subrefraction, 1435-36, 2559 subscriber identity module, 906 subscriber links, in satellite communications, 1251 subset construction, in constrained coding techniques for data storage, 575 subspace blind multiuser detection, 298, 301-302 subspace algorithm, in channel modeling, estimation, tracking, 406-407 substitution attack, 219, 222 substitution coding, 581 substitution method, in constrained coding techniques for data storage, 578 subtractive interference cancellation, in wireless multiuser communications systems, 1617 subvectors, in vector quantization, 2127 subword closed systems, 573 successive interference cancellation, 1617 sudden ionospheric disturbance, 949, 2066, 2067 suites of protocols, 541 sum product decoding, LDPC, 1309, 1312-15, 1314 Sunda's FSK, minimum shift keying, 1472 sunspot activity and radiowave propagation, 1477, 2061, 2063, 2065 super high frequency, 208 superfinger demodulator, 357 supergain antenna arrays, 160-161 superheterodyne receivers, 1478 superrefraction, microwave, 2559-60 supertrunks, community antenna TV, 512 supervisory frames, high level data link control, 546 suppressed carrier signal, amplitude modulation, 134 surf zone acoustic telemetry experiment, 24 surface acoustic wave, 327 surface acoustic wave filters, 2441-61 antenna duplexers and, 2458-59 applications for, 2457-60 chirp modulation and, 441, 447 clock recovery circuits and, 2460 convolvers and, 2455-56, 2455, 2456, 2460 finite impulse response filters and, 2450-52, 2456 piezoelectricity and, 2444-45 resonators and, 2454-55, 2455, 2456-57 surface equivalence principle, in antenna modeling, 172, 172 surface modeling, in antenna modeling, 175-176, 176 surface roughness (specular effects), radiowave propagation, 211–213, 212, 213 surface skimming bulk waves, 2441, 2444 surface transverse waves, 2441, 2444 surface waves, in leaky wave antenna, 1237 survivable optical Internet, 2461-72 sustained cell rate, 117, 266, 551, 1656, 1658 Svensson, Arne, 2274 swept power measurements, simulation, 2290 switch transmit diversity, 1586 switched beam antenna, 191-192 switched virtual circuits, ATM, 265-266 switches (see also optical cross connects and switches) Ethernet and, 1505-06, 1506 microelectromechanical systems and, 1354, 1355 switching, 549 acoustic echo cancellation and, 5 ATM and, 200-207, 200, 272-273 Banyan networks in, 202-203, 202 Batcher-banyan networks in, 203 buffering input and output in, 203-204, 203 bus matrix, 203-204 crossbar, 202-203, 202 distributed intelligent networks and, 719-29, 722, 726

taps, in community antenna TV, 518, 518

switching (continued) multistage interconnection networks in, 202, 202 NxN crossbar, 201-203, 202 routing and wavelength assignment in WDM and, 2098 satellite communications and, 2113 satellite onboard processing and, 482 space division, 202-203 switching diversity, radio resource management, 2093 switching modulator, amplitude modulation, 138-139, 138 symbol by symbol detectors, 1932 symbol duration, in modulation, 1335 symbol error probability, QAM, 2043, 2047-52, 2048, 2049, 2050 symbol error rate, Reed-Solomon coding for magnetic recording channels, 473 symbol fields, in BCH (nonbinary) and Reed-Solomon coding, 253 symbolic dynamic models, chaotic systems, 422 symbolic dynamics, chaotic systems, 427 symmetric ciphers, cryptography, 1152 symmetric key/private key encryption, 606, 607-609, 607 synapses, in neural networks, 1675-76, 1676 synchronization, 545-546, 2472-85 acoustic telemetry in, 23 additive white Gaussian noise, 2473-85 baseband and, 2479 carrier frequency and phase, 2481 cdma2000 and, 367 channel impulse response and, 2474-85 chaotic systems and, 422 code division multiple access, 2479-81, 2479 coding tracking in, 2480 coding, 2480-81 constrained coding techniques for data storage and, 576 continuous phase modulation and, 2473-85 digital audio/video broadcasting and, 2481 drive-response, chaotic systems and, 422, 422 feedforward/feedback systems in, 2475 frame, 2482-83 frequency and timing estimation in, 2482 frequency, 2475-77 global system for mobile and, 912-913 hard disk drives and, 1320 intersymbol interference in, 2474-85 mean square error and, 2475 minimum shift keying and, 1471-72, 1471, 1472 multicarrier transmission and, 2481-82 multiple access interference and, 2479-85 narrowband, 2473-85 orthogonal frequency division multiplexing and, 2481-85, 2481, 2482 phase shift keying, 2473-85 phase, 2477-78 pulse amplitude modulation and, 2028-30, 2029 pulse position modulation and, 2031 quadrature amplitude modulation and, 2052-57, 2473-85 RAKE receivers and, 2481 signal to noise ratio, 2473-85 spread spectrum and, 2479-81, 2479 timing, 2478-79, 2478 wideband CDMA and, 2878-79 synchronization channels, in IS95 cellular telephone standard, 349, 350 synchronous CDMA, 1096 synchronous connection oriented link, 313, 315 synchronous demodulator (see also phase coherent demodulator), 134 synchronous digital hierarchy, 2485-2510, 2486, 2485, 2509 administrative units in, 2497, 2498 ATM and, 201 bit rates in, 2496-97 containers in, 2498 dense WDM and, 748-757, 749 frame structure in, 2500

frequency justification in, 2506-2508

interconnection of STMs in, 2498-99 multiplexing in, 2498-2504, 2498, 2499, 2500, 2501-2506 optical cross connects/switches and, 1798 optical fiber and, 1798, 2615 pointers in, 2498, 2503-08, 2506-2508 scrambling in, 2499-2500 signal quality monitoring and, 2269 software radio and, 2307 synchronous transport module in, 2497-98 tributary units in, 2497-98 virtual containers in, 2497 wavelength division multiplexing and, 2863-73 synchronous optical network (see SONET) synchronous round robin with reservation, 1557 synchronous transmission, 545-546, 546, 1495, 1808 synchronous transport module, 2497-98 syndrome decoder, for cyclic coding, 620 syndrome equations for decoding, 247-248, 255, 623 synthesis, in antenna arrays, 153-157, 187 synthetic aperture radar, 1393, 1393 synthetic environment tactical integration visual torpedo program telemetry, 26, 27 system capacity, in radio resource management, 2090 system distance, in acoustic echo cancellation, 6 systematic (Cartesian) authentication coding, 220, 221 systematic coding cyclic coding and, 619, 619 Reed-Solomon coding for magnetic recording channels and, 469 sequential decoding of convolutional coding and, 2141 threshold coding and, 2579 t error correcting coding, 244, 253 T1 lines, 263 tabu lists, in quantization, 2130 tailbiting convolutional coding, 2511-16, 2513 a posteriori probability decoders in, 2515 applications for, 2516 BCJR algorithm for, 2515 decoding of, 2515-16 Viterbi algorithm for, 2515 tamed frequency modulation, 584-593 tank circuits, active antenna, 62 Tanner graph low density parity check coding and, 1311, 1311, 1312 product coding and, 2011, 2011 tap leakage algorithm, in adaptive equalizers, 87 tape drive, for Reed-Solomon coding for magnetic recording channels, ECC, 474 tapered microstrip transition waveguide, 1400-01, 1400 tapering, in leaky wave antenna, 1241, 1242 tapped delay line equalizers, 1688-96 baseband communications systems and, 1690, 1690 decision directed feedback equalizers and, 1693 decision feedback equalizers and, 1688-89, 1692 feedback filters in, 1690 feedforward filters in, 1690 filters and, 1690 finite length and, 1692-94, 1694 high definition TV and, 1689 impulse response and, 1689, 1689 infinite length symbol spaced equalizers for, 1690-92 intersymbol interference and, 1688 linear equalizers and, 1688, 1691–92 matched filters and, 1691 minimum mean square error detectors and, 1690, 1691, 1692 nonuniformly spaced, 1689-90, **1690** optical synchronous CDMA systems and, 1815, 1815, 1816 performance example of, 1694-95, 1695 quadrature amplitude modulation and, 1690 signal to noise ratio in, 1690 simulation and, 2289, 2289 sparse multipath channels and, 1688-96 tropospheric scatter communications and, 2700-02

zero forcing, 1690

history and development of, 2485-87

target lists, for multicasting, 1536 target probability, in satellite communications, 2119-20 Taylor distribution (Chebyshev error), 154-157, 155, 157. 187 Taylor one-parameter distribution, antenna arrays, 155 TCP for transactions, 2120 TCP friendly rate control, congestion control, 1662 TCP Reno (see also transmission control protocol), 1625, 1630, 1628, 1662 TCP segments, 541, 544 TCP Tahoe (see also transmission control protocol), 1625, 1628, 1662 TCP Vegas (see also transmission control protocol), 554, 1625, 1628-30, 1662 TCP/IP, 540-544, 1644 admission control and, 114 ATM and, 264, 273 broadband and, 2661-63 IP networks and, 267, 268 packet switched networks and, 1912 radio resource management and, 2094-95 satellite communications and, 1232-33, 2113, 2120 transmission control protocol and, 2603 virtual private networks and, 2807 TD/CDMÂ, 2589-90, 2592 TDSCDMA cell planning in wireless networks and, 369 cell planning in wireless networks and, 385-386, 385 teleconferencing, 540 Teledesic satellite, 196, 484 telegraph, 1477 telemetry (see also acoustic telemetry), 20-22, 21, 37 telephone acoustic echo cancellation and, 1 hands free, 1 IP telephony and, 1172-82, 1173 powerline communications vs., 1998 transducers (acoustic) and, 29 in underwater acoustic communications, 22-23, 36 wireless IP, 2931-41 wireless local loop standards and systems in, 2947-59, 2948 telephony and Internet protocol harmonizatioin over networks, 1181 telephony routing over IP, 1181 telescopes, free space optics, 1863-64 telesonar modems, 2215 Telesonar telemetry system, 24 television (TV), 1478 active antenna and, 50 antenna arrays and, 141, 180, 187, 2517-36 high definition (see high definition TV) satellite communications and, 880 terrestrial digital, 2547-55 trellis coded modulation and, ATSC and, 2631-32, 2632 Television Allocations Study Organization, 515 Telnet, 540, 541, 544, 1651, 2608 temporally ordered routing algorithm, 2211, 2888 10Base2, 1283 10Base5, 1283 10BaseT, 1283, **1283** 100BaseT, 1283-84 TeraNet, 1720 ternary sequences, 2536-47 autocorrelation and, 2541-42 bounds on sequence correlation in, 2543-44 code division multiple access, 2536-47 cross correlation in, 2542-43 Galois fields in, 2538-47 maximal length sequences (m sequences) in, 2540-41 phase shift keying and, 2536-47 sequence families with low correlation in, 2544-46 trace function in, 2539-40 terrain scatter and diffraction, in millimeter wave propagation, 1445 terrestrial digital TV, 2547-55 additive white Gaussian noise, 2547 advanced television systems committee and, 2549, 2550

terrestrial digital TV (continued) advantages and disadvantages, 2547-48 audio coding and, 2552-53 bit error rate in, 2547 channel coding in, 2548-49 digital video broadcasting and, 2549, 2550 diversified transmission in, 2554-55 indoor reception of, 2554 integrated services digital broadcasting and, 2549, 2551 - 52intersymbol interference and, 2548 mobile reception of, 2554 modulation in, 2549-50 MPEG compression and, 2552-53, 2553 NTSC standard and, 2546 orthogonal frequency division multiplexing and, 2549 PAL standard and, 2546 program and system information protocol in, 2553 quadrature amplitude modulation and, 2550-55 SECAM standard and, 2546 services and coverage for, 2553-55 signal to noise ratio, 2547 single frequency networks and, 2554-55 source coding for, 2552-53 standards for, 2546, 2550-52 transmission system for, 2548-52, 2549 transport layer and, 2553, 2553 video coding and formats, 2552 terrestrial microwave communications (see also microwave communications), 2555-72 test and measurement of optically-based high speed digital communications systems, 2572-79 **TETRA**, 379 text to speech systems, 1304-05, 2380-82 thermal asperity, in digital magnetic recording channel, 1326 thermal drift, in optical modulators, 1747 thermal noise adaptive antenna arrays and, 71-72, 74, 74 community antenna TV and, 514-515, 523-524 free space optics and, 1858 optical fiber systems and, 1843 parabolic and reflector antenna and, 1922, 1926-27 tropospheric scatter communications and, 2697, 2698 thermocapillary switches, 1792-93, 1792 thermoelectric element lasers, 1781 thermooptic planar lightwave circuits, 1703-04 thermooptic switches, 1785 Thevenin equivalent circuits, antenna, 184, 185 thin film filters (see also dielectric thin film stack interference filters), 1723, 2446 thinned arrays, 162-163 thinnet/cheapernet, 1283 thin-wire theory for antenna, 175 third-generation wireless, 125-126, 371, 385-391, 1479, 1483, **1483** microelectromechanical systems and, 1350 satellite communications and, 2115-16 session initiation protocol and, 2198 Third Generation Partnership Project, 126, 358, 397, 918, 1316, 2932 third order dispersion, solitons, 1766-67 third party call control, in session initiation protocol, 2198 threading, protocols, in MAC, 1348 3D antenna array, 151-152, 152 3G.324 standard, 2980 threshold coding/ decoding, 2579-85 a posteriori probability (APP) algorithm and, 2580, 2584 block coding and, 2583-84 convolutional coding and, 2581-83, 2582 convolutional self-orthogonal coding and, 2582-83, 2584 generator matrix in, 2579 Hamming coding and, 2579–80 history and development of, 2580 L step orthogonalization in, 2580 parity checks and, 2580-81 systematic encoding in, 2579 threshold current, in lasers, 1778

threshold sampling in waveform coding, 2837 thresholds, in speech coding/synthesis, 2364, 2364 throughput ALOHA protocols and, 342 automatic repeat request and, 226, 228-230 broadband and, 2655 cell planning in wireless networks and, 385, 385, 386 frequency division multiple access (frequency division multiple access and, 827, 827 media access control and, 1344 packet rate adaptive mobile receivers and, 1902-03, 1902 transmission control protocol and, 554 wireless packet data and, 2984-85 throughput per slot rate, ALOHA protocol, 128 Thuraya satellite onboard processing, 484 time congestion, in traffic engineering, 491, 491 time constrained ALOHA, 131 time delay, 1436, 1937 time deterministic time/wavelength division multiple access 1555 time difference of arrival, 2690, 2963 time diversity, 276, 371, 1481 time division CDMA, 123 time division coded modulation, unequal error protection coding, 2763-69, 2764, 2763 time division duplex adaptive receivers for spread-spectrum system and, <u>9</u>6 admission control and, 123 Bluetooth and, 310 cell planning in wireless networks and, 385-386 time division multiple access (time division multiple access, 458, 825, 2274, 2585-92 acoustic telemetry in, 25 adaptive receivers for spread-spectrum system and, 95-96, **96** admission control and, 120, 123 advanced mobile phone service and, 2586 antenna arrays and, 163 ATM and, 2907-09 automatic repeat request and, 225 Bluetooth and, 309-310 broadband wireless access and, 318, 320 burst types for, 2588, 2588 cable modems and, 324, 334-335 carrier sense multiple access and, 340-341, 340 cdma2000 and, 358 cell planning in wireless networks and, 377, 380 cellular telephony and, 1479 chann/in channel modeling, estimation, tracking, 409-410 chirp modulation and, 445, 446 cochannel intereference and, 448, 453-454, 453, 455 code division multiple access and, 2586, 2590 community antenna TV and, 523 enhanced data rate for GSM evolution and, 2589 frequency division multiple access (frequency division multiple access and, 828, 829, 2586 general packet radio service and, 2589 global system for mobile and, 911-912, 911, 2589 hybrid systems using, 2586, 2588 IMT2000 and, 1095-1108 intelligent transportation systems and, 503, 504-505 interference and, 1130-41 local multipoint distribution service and, 318, 320 media access control and, 1343-1347 mobile radio communications and, 1481-82, 1482 Nordic Telecommunications and, 2586 optical fiber and, 1808 orthogonal transmultiplexers and, 1880-85 packet rate adaptive mobile receivers and, 1886 physical layer subscriber signals in, 2586-89 polyphase sequences and, 1975-76 powerline communications and, 2003 principles of, 2586 radio resource management and, 2090, 2091-93 satellite communications and, 878-881, 1231-32, 1231, 1253 satellite onboard processing and, 477, 479

threshold effect, 820-821

shallow water acoustic networks and, 2208, 2212, 2215 signal and system structure in, 2586-89, 2587 software radio and, 2312, 2312, 2314 spatiotemporal signal processing and, 2336 system structure for, 2589, 2589 TD/CDMA and, 2589-90 turbo product coding and, 2727-37 underw/in underwater acoustic communications, 44 wireless local loop and, 2950-51, 2955 wireless multiuser communications systems and, 1602, 1609 time division multiplex, 1906 ALOHA protocol and, 130 Bluetooth and, 315 broadband wireless access and, 318, 320 Ethernet and, 1512 flow control and, 1625 H.324 standard and, 920-922, 921 local multipoint distribution service and, 318, 320 medium access control and, 1552, 1553 multibeam phased arrays and, 1514 optical cross connects/switches and, 1799 optical modulators and, 1741 tropospheric scatter communications and, 2693 wavelength division multiplexing and, 2864 weighted, 1552 time division WDMA, 1553 time domain, in antenna modeling, 169 time domain constraints, in optical recording, 579 time invariant digital filters, 687, 700 time of arrival, in wireless systems, 2690 time representation (TSR) sequence, 288 time reversal space time block coding, 2329, 2330 time slot assignment, in ADSL, 1574, 1575 time slots, 1625, 1667 time spread multiple access, 1348 time stretching using dispersive optical elements, 1968, 1968 time to live field, MPLS, 1594-95 time variance in radio channels, 393, 394-395 time varying coding, 581 time varying maximum transition run length coding, 1332 timeouts and retransmission, in TCP, 2611-12 timescale, 1667, 1908 timing and synchronization, 2478-79 timing errors, orthogonal frequency division multiplexing, 1875–76 timing jitter, 1759, 1767 timing window, in constrained coding techniques for data storage, 573 tinygrams, 2609 TM multicasting algorithmd, 1533 Toeplitz symmetric matrix, in linear predictive coding, 1263 token bus, 547 token ring, 345, 547, 549, 1345-46, 1529 tonpilz sonar transducer, 35, 35 top down processing, in speech coding/synthesis and, 2363 topologies ad hoc wireless networks and, 2886 Ethernet and, 1505, 1505 fault tolerance and, 1632-33 free space optics and, 1851 optical fiber and, 1715-17 powerline communications and, 1999-2000, 1999 reliability and, 1632-33 routing and wavelength assignment in WDM and, 2100-01 shallow water acoustic networks and, 2208 wavelength division multiplexing and, 651 wireless communications, wireless LAN and, 1285, 1285 trace function, in feedback shift registers (FSR), 796 track density, in hard disk drives, 1321 tracking channels (see channel tracking in wireless systems) satellite communications and, 1252 ultrawideband radio and, 2761

tracking and data relay satellite, 1519 tracks, in storage media, 1320-21, 2248 traffic activity factors, 1608 traffic analysis attack, 1646 traffic channels, in IS95 cellular telephone standard, 349 traffic computation, in cell planning in wireless networks, 380 traffic conditioning, 1660 traffic conditioning agreement, DiffServ, 270 traffic contracts, ATM, 205 traffic coverage rate, cell planning in wireless networks, 374 traffic engineering (see also flow control), 485-501, 549, 2462 A erlangs in, 488 ALOHA protocols and, 499-501, 499, 500 arrival process in, 486-491 arrival statistics in, 486 arrivals in, 489 ATM and, 273 blocked calls delayed in, 495-497 blocked calls held in, 497-499 blocking in, 486-487, 486 blocking probability in, 498 busy hour in, 487-488 call congestion in, 491, **491** calls in, 485 capacity and, 492, 494 cell design and, 490 centum call seconds in, 488 channels in, 485 collisions in, 500 congestion and, 491, 491 day to day variation in, 488 distribution of arrivals in, 489 economies of scale in, 494 Erlang B blocking in, 487, 491-499 Erlang C blocking and, 487, 495-497, 495, 498 Erlangs in, 486 examples of, 486-487 first in first out models in, 487, 495 flow conservation principle in, 493-494, 493 flow control, traffic management and, 1654 holding time in, 485 infinitesimal generator model for, 488-489 interarrival times in, 489 load defined for, 488 Markov chains in, 492 merging Poisson processes in, 489 multiprotocol label switching and, 271, 1599 network design and, 501 offered load in, 486 packet switched networks and, 500 packet time and, 500 peakedness in, 490-491 Poisson arrival process in, 488-491, 491, 497-499, 498 queuing probability in, 496 random access protocols for, 499-501, 499 random arrivals in, 489 random service order and, 498 seizure of a server by call in, 492 service facility in, 485, 485, 492 splitting Poisson processes in, 489-499, 499 state space in, 492-493, 493 state transition diagrams in, 492-493, 493, 495, 495, 498 stationarity in, 487 three-node network example for, 490, 490 time congestion in, 491, 491 trunking efficiencies in, 494 utilization rates in, 486-487, 486 virtual private networks and, 2809 traffic management (see also flow control), 1653-65 active queue management in, 1661 adaptive virtual queue in, 1661 additive increase multiplicative decrease in, 1662 admission control and, 1655-56 asynchronous transfer mode and, 1654, 1656, 1658 ATM and, 205-206, 266 buffer management and, 1654, 1656, 1659-61

common open policy service in, 1656 congestion control and, 1653, 1661-63 connection level controls in, 1654 connections in, 1653 constraint-based routing in, 1654-55 delay in, 1660 differentiated services and, 1654, 1657-60, 1658 dropping in, 1659 DropTail in, 1660 earliest due date algorithm in, 1660 early packet discard in, 1661 explicit congestion notification in, 1662-63, 1662 explicit rate feedback in, 1663 explicit rate indication for congestion avoidance in, 1663 fairness in, 1653 first in first out in, 1660 flows in, 1653 forward acknowledgement in, 1662 generalized processor sharing in, 1660 generic cell rate algorithm in, 1656, 1659 integrated services and, 1654, 1657, 1657 intermediate system to intermediate system in, 1658 internet gateway protocols in, 1658 Internet protocol and, 1653 interoperability among different architectures in, 1658–59 jitter in, 1660 leaky bucket algorithm in, 1659 lightweight directory access protocol in, 1656, 1656 macroflows in, 1653, 1654 marking or tagging, 1659 measurement-based admission control in, 1656 microflow in, 1653 multiprotocol label switching and, 1658, 1658, 1659 objectives of, 1653 open shortest path first in, 1658 packet classifiers in, 1656 packet schedulers in, 1656 partial packet discard in, 1660-61 policy control, 1656 policy decision points in, 1656 policy enforcement points in, 1656 policy routing and, 1654-55 proportional integrator in, 1661 quality of service and, 1653, 1654, 1655 queues in, 1661 queuing delay in, 1653 random early detection in, 1661, 1661 random early marking in, 1661 real time control protocol in, 1662 reservation protocols and, 1655, 1656-57 resource allocation/utilization in, 1653, 1663, 1663 resource reservation protocol and, 1655-59, 1657 scheduling in, 1654, 1660 selective acknowledgement in, 1662 self-clocked fair queuing in, 1660 session in. 1653 session level controls in, 1654 simplicity and Occam's razor concept in, 1653 stability and, 1653 stop and go algorithm in, 1660 TCP friendly rate control in, 1662 TCP Reno in (see also transmission control protocol), 1662 TCP Tahoe in (see also transmission control protocol), 1662 TCP Vegas in (see also transmission control protocol), 1662 traffic characteristics in, 1656 traffic conditioning in, 1660 traffic engineering and, 1654 traffic policing in, 1659-60 traffic shaping in, 1654 transmission control protocol and, 1653, 1661-62 trunking in, 1654 "unhappiest source" concept in, 1653 usage parameter control and, 1659 user datagram protocol and, 1653, 1662 user network interface and, 1657 virtual clock in, 1660

weighted fair queuing in, 1660 worst case fair weighted fair queuing in, 1660 traffic modeled admission control, 117 traffic modeling, 1666-75 asynchronous transfer mode and, 1672 autocorrelation in, 1667, 1669, 1669 autoregressive processes in, 1666, 1668 bandwidth and, 1666 batch arrivals in, 1667 burstiness of traffic and, 1666, 1667, 1668 concepts of, 1666-67 data source models in, 1671 decay in. 1667 discrete autoregressive model in, 1666, 1668 discrete time models in, 1667 distorted Gaussian models in, 1670 effective bandwidth and, 1671, 1673 elastic sources and, 1671, 1672-73 envelope processes in, 1666, 1673 fluid models in, 1670-71 fractal Brownian motion models in, 1669-70 fractal Gaussian noise models in, 1669-70 fractal Levy motion models in, 1670 heavy tailed on/off models in, 1669 local scaling components in, 1670 long range dependent models in, 1667, 1668-70 Markov/semi–Markov models in, 1666, 1667–68 matching and fitting in, 1666 monofractal models in, 1669 MPEG compression and, 1672 multifractal models in, 1670 probability density function in, 1667 quality of service and, 1673 randomness of traffic and, 1666 renewal models in, 1666, 1667 round trip time in, 1672 self-similar processes in, 1669 short range dependent models in, 1667-68 simple vs. compound traffic in, 1667 source models in, 1671-73, 1671 sources of traffic and, issues concernig, 1666 spatial renewal process model in, 1666, 1668 streaming sources and, 1671 time slots in, 1667 timescale in, 1667 transform expand sample processes in, 1666, 1668 transmission control protocol in, 1671, 1672 video source models in, 1671-72, 1672 wide sense stationary in, 1667 workload and, 1666, 1667 traffic padding, 1646, 1649 traffic policing, 1659-60 traffic shaping, 1, 551-552, 1563-64, 1654 traffic verification or policing algorithms, 1563 training automatic speech recognition and, 2384, 2387 blind equalizers and, 287 in channel modeling, estimation, tracking, 398, 401-403, 409 equalizers and, 82, 286 multiple input/multiple output systems and, 1453 neural networks and, 1675, 1677-79, 1677 quantization and, 2129 spatiotemporal signal processing and, 2338 wireless multiuser communications systems and, 1614-15 training sets, vector quantization, 2125 training signals, in adaptive receivers for spread-spectrum system, 100 training systems, in spatiotemporal signal processing, 2333 transactions, in session initiation protocol (SIP), 2198 transceivers multiple antenna, for wireless, 1579-90, 1580 wireless multiuser communications systems and, 1608-20 transcoding, in image and video coding, 1057 transducers, acoustic (see acoustic transducers) transduction, 34 transfer electron devices, 51 transfer function, in neural networks, 1676

transfer functions, in digital filters, 690, 691 transform coding, 2593-2603, 2594 analysis transform T in, 2601 bit allocation in, 2597-98 centroids in, 2597 compression and, 646-648, 645 constrained source coding in, 2594-96, 2595 departures from standard model in, 2601-02, 2602 easy optimization in, 2600 entropy coding and, 2596 history and development of, 2602 Karhunen-Loeve transforms in, 2599-2600, 2599 mean square error and, 2593 model for, 2596-98 optimal transforms in, 2599-2601 partition cells in, 2599 probability density function in, 2597 quantization in, 2594, 2597 quantizers and, 2596-97 reproduction codebook in, 2596 scalar and vector coding and, 2601-02, 2602 source coding and, 2593-94, 2594 synthesis transform U in, 2600-01 unconstrained source coding and, 2593, 2595 visualization of, 2599-2600 waveform coding and, 2836-37, 2837 transform domain analysis, in digital filters, 690-691 transform expand sample processes, in traffic modeling, 1666, 1668 transformation matrix, in adaptive antenna arrays, 70-71 transforms, vector quantization, 2125-26 transient response, 851-853, 852, 853, 851 transimpedance, in optical transceivers, 1833 transistors, 51, 65, 262 transition shift, in digital magnetic recording channel, 1325, 1325 translating repeater, in satellite onboard processing, 476, 476 transmission control protocol, 268, 541, 543-544, 544, 2603 broadband and, 2661-62, 2662 congestion control and, 553-554, 1661-62, 2610-11, 2611 connection establishment/termination in, 2607-08, 2607, 2607 data transfer in, 2608-16 differentiated services and, 672-673, 673 fault tolerance and, 1632, 1640 flow control and, 1625, 1627-28, 1653, 1661-62 future of, 2612 headers in, 2605-06, 2605 hypertext transfer protocol and, 2604 Karn's algorithm and, 2612 maximum segment size in, 2604, 2606 message format in, 2605–07, **2605**, **2606** multimedia networks and, 1566 multiplexing and, 2604 Nagle algorithm in, 2609 operation of, 2607-12 optical fiber and, 2618-19 options for settings in, 2606-07 packet switched networks and, 1911, 1912 piggybacking in, 2608 reliability and, 1632, 1640 rlogin and, 2608 round trip time in, 554, 2609, 2612 satellite communications and, 1233, 2120 security and, 1646 segments in, 2604 session initiation protocol and, 2197 sliding window mechanisms in, 2604, 2609-11, 2609 streaming video and, 2439 TCP Vegas implementation in, 554 Telnet and, 2608 throughput in, 554 timeouts and retransmission in, 2611-12 traffic modeling and, 1671, 1672 transport protocol data units in, 2604 user datagram protocol and, 2604 wireless IP telephony and, 2934 wireless packet data and, 2984, 2989

transmission line models, in microstrip patch antenna, 1357 transmission lines active antenna and, 50-52, 50, 54-55 Ethernet and, 1506-07 lossy vs. lossless, 55-56, 56 microelectromechanical systems and, 1353-54 transmission loss (see also losses), 15, 2067 transmission paths, in millimeter wave propagation, 1443-48 transmitted far field method, for optical fiber, 435 transmitted near field method, for optical fiber, 435 transmitter power control, 1982-88 transmitter scheduling algorithm, 1557 transmitter/receiver network, in acoustic modems for underwater communications, 17 transmitters acoustic telemetry in, 23, 23 active antenna and, 50-51, 51 continuous phase frequency shift keying and, 594, continuous phase modulation and, 590, 590 digital audio broadcasting and, 681, 681 discrete multitone and, 740, 742 free space optics and, 1851-52, 1852 holographic memory/optical storage and, 2135–36, 2135 local multipoint distribution services and, 1268 microwave and, 2567-70, 2568 minimum shift keying and, 1462-67, 1462, 1465, 1467 multicarrier CDMA and, 1522-25, 1522, 1524, 1525 multiple input/multiple output systems and, 1450-56, 1450 optical (see also optical transceivers), 1824-40 optical communications systems and, 1484, 1488-91, 1489, 1491 optical fiber and, 1707 optical memories and, 1733 optical synchronous CDMA systems and, 1815-16, 1816 orthogonal frequency division multiplexing and, 1867-71, 1869 power control in, 1982-88 pulse amplitude modulation and, 2022-30 pulse position modulation and, 2031-32 satellite communications and, 878, 2111 signature sequence for CDMA and, 2275-76, 2275 spark gap, 48-49, 48 terrestrial digital TV and, 2548-52, 2549 tropospheric scatter communications and, 2695-97 ultrawideband radio and, 2757, 2757 in underwater acoustic communications, 42-43 wavelength division multiplexing and, 651 wireless infrared communications and, 2926 wireless multiuser communications systems and, 1608-20 transmultiplexers, orthogonal, 1880-85 transparency, 2269, 2864, 2867-71 transparency of protocols fiber networks, 1797 transponder, satellite, 476, 878 transport layer OSI reference model, 540 packet switched networks and, 1911 streaming video and, 2436-37 TCP/IP model, 541 terrestrial digital TV and, 2553, 2553 transport layer security, 1154-55, 2202 transport protocol data units, 2604 transport protocols for optical networks, 2513-22 asynchronous transfer mode and, 2619-20 congestion control and, 2616-17 flow control and, 2616-17 framing, segmentation, reassembly in, 2617 implementation issues for, 2620-21 in band vs. out of band signaling, 2618 interrupt coalescing in, 2620 memory and, 2620 multicast transfers in, 2615 multiplexing and, 2617-18 NETBLT and, 2616

reliability in, 2615-16 service specific connection oriented protocol and, 2616, 2619-20 state management and, 2618 transmission control protocol as, 2618-19 unicast transfers in, 2615 VMTP and, 2616 Xpress transport protocol and, 2616, 2617, 2620 transverse electric modes, in waveguide, 1393-96 transverse electromagnetic waves, 182 transverse magnetic modes, in waveguide, 1393 trap door one way functions, 606, 609-610 traveling wave antenna, 1258-60, 1259 traveling wave tube amplifiers, 531, 532, 535, 1457 traveling wave tubes, 2567 traveling waves, in waveguide, 1394 tree algorithms, in media access control, 1347 tree structured search, 644, 2126, 2129 trellis coded modulation, 590, 2622-35 additive white Gaussian noise, 2623-24, 2624, 2629 applications for, 2631-342 asymptotic coding gain in, 2628 automatic repeat request and, 225 bit error probability in in, 2629 bit error rate in, 2629 bit interleaved coded modulation and, 276–286, **276** block coding and, 2622 cable modems and, 330 decoding in, 2627 design principles in, 2625-29 edge in, diverging and merging, 2627 Euclidean distance in, 2527-29, 2622, 2627 fading and, 2633-34 free space distance in, 2527-29 Gray coding and, 2625 Hamming distance and, 2634 minimum shift keying and, 1469-70, 1470 modems and, 1497, 1498 multidimensional, 2633 multiple, 2634 performance of, 2629-30 phase shift keying and, 2622-35 power spectrum for, 2631 pulse amplitude modulation and, 2625 quadrature amplitude modulation and, 2624-35 quadrature phase shift keying and, 2622-35 redundancy in, 2623 rotational invariance in, 2632-33 satellite transmission and, 2631 set partitioning in, 2625-26, 2626 Shannon or channel capacity in, 2623 signal to noise ratio, 2623-24, 2628, 2630, 2634 tables of coding for, 2630-31 television and, 2631–32, 2632 trellis coding and, 2637-53 trellis construction in, 2626-27, 2626 turbo coding and, 2737-53 V.32 modems and, 2632, 2632, 2633 Viterbi algorithm and, 2627 wireless and, 2922 wireless multiuser communications systems and, 1610 trellis coding, 556, 2635-53 a posteriori probability algorithm in, 2648, 2650 additive white Gaussian noise, 2636, 2648 automatic repeat request and, 2635 bandwidth in, 2636-37 BCH coding and, 2640 binary phase shift keying in, 2638-53 bit error rate in, 2637, 2638, 2639 bit interleaved coded modulation and, 279-280, 279, 282 - 283branch metrics in, 2647 community antenna TV and, 526 compression and, 644-646, 645 construction of, 2642-45 continuous phase modulation and, 587 convolutional coding and, 2645-46, 2645 decoding in, 2646-49 encoding in, 2635 equalizers and, 91, 91

metric calculation for first stage decoding in,

trellis coding (continued) error correction coding and, 2635, 2639-40 error detection and correction in, 2639-40 Euclidean distance in, 2642 finite state machine in, 2640-42, 2641 forward error correction and, 2635-40 hard vs. soft decision in, 2640 iterative coding and, 2640, 2650-51 iterative decoding and, 560 magnetic recording systems and, 2260-61, 2260 magnetic storage and, 1331-33 maximum likelihood decoding in, 2638, 2646-48 modulation in, 2635 multidimensional coding and, 1538 multiple input/multiple output systems and, 1455 parallel concatenated (see turbo coding) power control in, 2636-37 quadrature amplitude modulation and, 2636-53 quadrature phase shift keying and, 2637-53 Reed-Solomon coding and, 2640 sequence decoding in, 2646-48 sequential decoding of convolutional coding and, 2142, 2144, 2154-56 serially concatenated coding and, 2164, 2180 Shannon or channel capacity in, 2636-37 signal to noise ratio, 2637-39, 2642 single channel per carrier in, 2638 soft output decoding algorithms and, 2296, 2296 space-time coding and, 2325-26, 2328-29 spatiotemporal signal processing and, 2333 symbol decoding in, 2648-49 tailbiting convolutional coding and, 2513-16, 2513 terminating, 2296 trellis coded modulation and, 2637-53 truncation length or decision depth in, 2648 turbo coding and, 2637, 2640, 2649-51, 2650 in underwater acoustic communications, 45 V.34 modems and, 2640 Viterbi algorithm in, 2647, 2648 trellis diagrams, 600, 600, 601, 1618, 1618 trends in broadband communication networks, 2653-77 trends in wireless indoor networks, 2677-92 triangular inequality elimination, in vector quantization, 2126 tributary units, in synchronous digital hierarchy, 2497-98 triggered backup, 1634-35 triple transit echo, in surface acoustic wave filters, 2449 trivial file transfer protocol, 335 tropical anomalies in radiowave propagation, 2065 tropospheric propagation, 2059 tropospheric radiowave propagation, 208, 1477, 2059 tropospheric scatter communication, 215, 2692-2704 adaptive equalizers in, 2700-02 analog system performance using, 2697-98 decision feedback equalizer and, 2701-02, 2701, 2702 digital systems using, 2689-2703 diversity and, 2693-95 Doppler effect, Doppler spread in, 2698-99 error detection and correction in, 2701-02 fading in, 2698-99 frequency division multiplexing and, 2693 intermodulation noise, 2697, 2698 intersymbol interference and, 2699 losses in, 2695-96 matched filters in, 2700, 2700 maximum likelihood detectors and, 2702-03 multipath in, 2697-98 quadrature phase shift keying and, 2693, 2700 receivers for, 2699-2703 signal to interference plus noise ratio in, 2697 signal to noise ratio, 2697 tapped delay lines and, 2700-02 thermal noise and, 2697, 2698 time division multiplexing and, 2693 transmitters for, 2695-97 variability in, 2696 true time delay shifter, in microelectromechanical systems, 1355, 1355

truncated binary exponential backoff, Ethernet, 1281

trunking, in flow control, 494, 1654 trunking theory, 1478 trust, in authentication coding, 222-224 Trust software radio, 2305, 2316 trusted authority, authentication, 613-614 trusted entities, 1649 trusted third party, authentication, 613-614 truth tables, feedback shift registers, 790-791, 790, 793, 794 tunable bandpass filters, 1764 tunable lasers, 1780-81 Tundra orbit, 1250 tunneling general packet radio service and, 867 IP networks, 273 multiprotocol label switching and, 271, 1596-97 security and, 1651 virtual private networks and, 2809, 2811-12, 2812 turbo coding (see also concatenated convolutional coding and iterative decoding), 556-570, 2179, 2180, 2704-16 a posteriori probability decoders in, 2705, 2714 additive white Gaussian noise, 2703 applications for, 2714-15 binary phase shift keying and, 2704-16 bit error rate and, 2704–16 bit interleaved coded modulation and, 285 cdma2000 and, 367 circular recursive systematic convolutional coding and, 2709, 2709 convolutional coding and, 600 decoding in, 2713-14, 2713 encoding/decoding in, 2705, 2705 error correction coding and, 2704-05 interleaving in, 1141-51, 1142-1149 iterative decoding and, 560 iterative detection algorithms for, 1182-96 logarithmic likelihood ratio and, 2713, 2714 low density parity check coding and, 1312, 1316 low density parity check coding and, 658 magnetic recording systems and, 2266-67 magnetic storage and, 1326-27 maximum a priori decoders and, 2705, 2714 multidimensional coding and, 1548-49, 1549 multiple input/multiple output systems and, 1456 parallel concatenated convolutional coding, 2710 permutation function in, 2711 product coding and, 2011, 2012, 2727-37 quadrature phase shift keying and, 2704-16 recursive systematic convolution (coding and, 2705-07, 2706 Reed-Solomon coding and, 2703 satellite communications and, 1229-30, 1230 serially concatenated coding and, 2164, 2176-77, 2176 soft input/soft output systems in, 2713, 2714 soft output decoding algorithms and, 2295, 2302, 2303 space-time coding and, 2328-29 termination of, 2708-09, 2708 trellis coded modulation and, 2737-53 trellis coding and, 2637, 2640, 2649-51, 2650 unequal error protection coding and, 2766–67 Viterbi algorithm and, 2705, 2714 wireless multiuser communications systems and, 1604, 1610 turbo equalization (see also turbo coding), 28, 2716-27 error correction coding and, 2716-27 intersymbol interference and, 2716-27 signal to noise ratio, 2716-27 soft input/soft output systems and, 2716-27, 2715 turbo trellis coded modulation, 2737-53 Bahl-Jelinek algorithm in, 2738 binary phase shift keying and, 2738-53 bit error rate and, 2738, 2747-49 code constraints in, 2740-42 component code search in, 2742 decoding and, 2738, 2743-47, 2744, 2745, 2746 encoding in, 2738-42, 2739, 2740 examples and simulations using, 2747-49 interleaving in, 2740-42, 2741 maximum a priori decoder in, 2743-45, 2750

2745-46 noise and, 2738 quadrature amplitude modulation and, 2738-53 quadrature phase shift keying and, 2738-53 signal to noise ratio, 2738 subset decoding in, 2747 Ungerboeck coding and, 2738 Viterbi algorithm and, 2738, 2743 turbulence, atmospheric (see also scintillations), 1861-63, 1861 TV SAT horn feed, 1392, 1392 two dimensional burst coding, multidimensional coding, 1538 two dimensional constraints, constrained coding techniques for data storage, 582 two dimensional dot coding, multidimensional coding, 1538 two layer spreading, code division multiple access, 2875-76, 2875 two stage conjugate algorithm, quadrature amplitude modulation, 2056 2n prime coding, optical synchronous CDMA systems, 1811-12 two-tone products, optical fiber, 1687 u law companders, 529, 529 U. S. Navy Advanced System Technology Office, 24 U.S. digital cellular systems, 1479, 1481 U.S. Naval Command Control and Ocean Surveillance Center, 24 Udaltsov-Shlyuger skywave method, 2062, 2064 ultra high frequency, 208, 1478, 2517-36 ultrasound, surface acoustic wave filters, 2441-43, 2441 ultrawideband (UWB) radio, 2754-62 additive white Gaussian noise, 2756, 2757 applications for, 2760-62 bandwidth and, 2761-62 binary phase shift keying and, 2755-62 code division multiple access, 2754-62 differential phase shift keying and, 2755 direct sequence, 2757 encoding and decoding in, 2755 frequencies for, 2754 impulses and impulse streams in, 2754–55 in building propagation of, 2758-59 link budget for, 2760 monocycle pulses in, 2755, 2755 multipath and, 2761 pseudonoise coding and, 2754 pulson application demonstration and, 2758 radar and, 2761 RAKE receivers and, 2757-59 receivers for, 2757, 2757 reception of impulses in, 2760 reflection and, 2759-60 Shannon or channel capacity in, 2760-61 signal propagation in, 2758-60 signal to noise ratio, 2756 spectrum allocation and, 2761-62 tracking in, 2761 transmitters for, 2757, 2757 wireless systems and, 2761 UMTS terrestrial access radio network, 2116 unbalance feed modified batwing antenna, 2517-36 uncompressed digital video, image and video coding, 1027 unconditionally secure authentication coding (see also authentication coding), 218 unconstrained source coding, transform coding, 2593, 2595 undersampling, photonic analog to digital conversion, 1960, 1961 underspreading, 395, 1604 underwater acoustic communications, 36-47 applications for, 36 axis of the deep sound channel in, 38 bandwidth in, 36-38 bit error rates in, 37 channel characteristics in, 37-40 control signals in, 37

underwater acoustic communications (continued) convergence zone in, 38 data compression for, 36, 37 data rates in, 37 digital signal processing in, 36, 43-44 Doppler effect and, 39-40, 39 encoding in, 43, 45-46 equalization in, 42-43 frequency division multiple access in, 44 history of, 36 interference cancellation, 44 intersymbol interference (ISI) in, 38, 44 loss in, 37-38, 38 mobile communications and, 46 modulation schemes in, noncoherent vs. coherent, 40-41, 45-46 multipath interference in, 38 multiuser networks for, 44-45 noise in, 37 optical communications vs., 36 phase coherent detection in, 41-43, 42 range in, 37-38 research topics in, 43-46 self-optimization in, 45 shallow water networks and, 2206-21 signal processing methods in, for multipath compensation, 41-43 signal to noise ratio in, 37-38 single sideband in, 36 speech signals in, 37 system design for, 37, 40-41 telemetry and, 37 telephone, 36 time division multiple access in, 44 time variation, delay, scattering, and fading in, 38-40, 39, 45-46 transmitters and receivers for, 42-45, 42 video signals in, 37 underwater acoustic telemetry (see also acoustic telemetry), 22-29 underwater communications (see acoustic modems for underwater communications underwater range data communications, 26-27 underwater telephone, 22-23 undetected word error, sequential decoding of convolutional coding, 2159 undisturbed error, acoustic echo cancellation, 6, 10 unequal error protection, 2355, 2762-69 unfairness factor (see also fairness), 2103 Ungerboeck coding, 2738 Ungerboeck set partitioning, 276, 276, 280-281 "unhappiest source" concept, 1653 unicast ing, 654–655, 1530, 2615 unidirectional path protection/path switched rings, SONET, 2495 unidirectional transducers, surface acoustic wave filters, 2449-50, 2449 unified modeling language, 2304, 2312 unified theory of diffraction, 216 uniform linear antenna arrays, 144-145, 146, 153, 153 uniform linear arrays, cochannel interference, 455 uniform linear virtual array, 69–71, 75, 77 uniform resource locators (URL), wireless application protocol, 2901 uniform theory of diffraction, path loss, 1936, 1942 union bounds, convolutional coding, 598-599, 605 universal approximators, neural networks, 1679 universal asynchronous receiver/transmitter, 1495 Universal Disc Format, 1736 universal mobile telecommunications systems, 371-372, 2963 admission control and, 120, 126 bit error rate in, 387 cell planning in wireless networks and, 384, 385-391 cellular communications channels and, 397 coding puncture rates in, 386 dynamic mode in, 388, 390-391 forward error control in, 386, 387 frequency division multiple access and, 828 general packet radio service and, 866

global system for mobile and, 907

H.324 standard for, 918-929, 919 Hadamard coding and, 929 IMT2000 and, 1096-1108 intelligent transportation systems and, 507 mobility portals and, 2191, 2194 network design in, 388-391, 389 orthogonal variable spreading factors in, 387 planning strategy for, 387 quality of service and, 387 quasidynamic mode in, 388, 389-390 RAKE receivers and, 387, 388 satellite communications and, 2116 service types in, 386-387 signal to interference ratio in, 386, 387 spreading sequences in, 386 static mode analysis in, 388, 389 wideband CDMA and, 2873-74 wireless IP telephony and, 2932-41 universal phone numbers, paging and registration, 1917–18 universal plug and play, 2194, 2310 universal resource identifier, session initiation protocol, 2196, 2198 universal theory of diffraction, indoor propagation models, 2018 universal wireless communication (UWC), wireless local loop, 2954 Universal Wireless Communications Consortium, 126 UNIX, 268 unlicensed bands, 308-309, 2678-79 unlicensed national information infrastructure, 2941, 2944 unmanned airborne vehicles, 169 unmanned undersea vehicles, 24, 28, 36 unnumbered frames, high level data link control, 546 unreliable protocols, 542 unshielded twisted pair, Ethernet, 1283, 1506-07 unslotted ALOHA, 128-132, 342-343, 343 unspecified bit rate, 123, 206, 267, 551, 1658 unsupervised learning, in neural networks, 1678 uplinks, satellite communications, 877, 1223, 1223, 2115 upper layer (see layer 3) signaling, 365 upper sideband, amplitude modulation, 133 upstream channel descriptor, cable modems, 334–335 usage parameter control, 205, 206, 1659 user agents, session initiation protocol, 2196, 2198, 2201 user datagram protocol, 544, 549, 2902 broadband and, 2662 congestion control and, 1662 flow control, traffic management and, 1653, 1662 packet switched networks and, 1911 security and, 1646, 1651 session initiation protocol and, 2197 streaming video and, 2439 transmission control protocol and, 2604 user network interface, ATM, 264, 265, 272, 1657 user plane, ATM, 264 user to network interface, 113-114, 1799 utility acoustic modem, 24 utilization rates, traffic engineering, 486-487, 486 UTRAN, 2873-74 V.32 modems, trellis coded modulation, 2632, 2632, 2633 V.34 modems, trellis coding, 2640 V.42bis compression, 1496 V.90 modems, 2770-79, 2771 validated queue algorithm, medium access control, 1556 validation, authentication coding, 219 Van Allen radiation belt, 1249, 2112 varactors, microelectromechanical systems, 1353, 1354 variable bit rate (VBR), 123, 267, 1563, 1566, 1567, 2420 variable decision circuits, signal quality monitoring, 2269-70, 2270 variable gain amplifier, magnetic recording systems, 2250 variable length coding, 1030, 2123 variable length message transmission, medium access control, 1551-55

variable length principal state coding, constrained coding techniques for data storage, 578 variable message signs, 505 variable optical attenuators, 1704 variable rate CELP, 2826 variance, 1339, 2129 VASCO project in intelligent transportation systems, 503 vector potentials and loop antennas, 1290-91, 1294 vector quantization, 2122-32, 2125-28, 2125 adaptive, 2128 classified, 2127 codebooks in, 2125 complexity barrier in, 2126 compression and, 642-644 discrete cosine transform in, 2125-26 encoding and decoding in, 2125 entropy constrained, 2128 exact vs. approximate methods in, 2126 finite state, 2127 image and video coding and, 1030, 1035-37, 1036 image compression and, 1065 L stage, 2127 lattice, 2127-28 mean distance ordered partial search in, 2126 mean removed, 2127 nearest neighbor problem in, 2126 open-, closed-, and semi-closed loop, 2127 partial distortion search in, 2126 prediction error or residual in, 2127 prediction techniques in, 2127 predictive vector quantization in, 2127 product codevector in, 2126 product coding and, 2127 quality and resolution in, 2126 reproduction vectors in, 2125 residual or multistage, 2127 shape-gain, 2127 speech coding/synthesis and, 2350, 2372 structures in, 2122-28 subvectors in, 2127 training sets in, 2125 transform coding and, 2601-02, 2602 transforms in, 2125-26 tree structured, 2129 tree structured search in, 2126 triangular inequality elimination in, 2126 unstructured vs. structured, 2126 Walsh-Hadamard transform in, 2126 waveform coding and, 2833-34 vector sum excited linear pulse, 2349, 2356, 2821, 2825, 2825 vector transformations, image and video coding, 1041 vectors, BCH coding, binary,, 239-240 Vernier effect, 1725, 1780 vertex, parabolic and reflector antennas, 1920, 2083 vertical Bell Labs layered space time scheme, 1587-89, 1589 vertical cavity surface emitting lasers, 1739, 1781, 1853 vertical plane launch approximation, 1942-43 very high frequency, 208, 1478, 2517-36 very high speed DSL, 2779-2807, 2781 architecture for, 2784-91 broadband wireless access and, 317 carrierless amplitude and phase and, 2791, 2801 cyclic suffix in, 2795-96, 2796, 2797 decision feedback equalizer, 2802, **2803** digital duplexing in, 2793 digital subscriber line access multiplexer and, 2784 digital TV and, 2780 discrete multitone and, 2791-2801, 2792, 2794 Ethernet in the first mile and, 2803-05, 2804 extension to, 2781-84 framing in, 2800-01 frequency division duplexing and, 2801 history and development of, 2779-84 home phone network of America and, 2790 initialization in, 2801 integrated services digital network and, 2780 modems for, 2779 multicarrier concept in, 2791-93

wireless multiuser communications systems and,

very high speed DSL (continued) multiple input/multiple output systems and, 2803-05 near end and far end crossfalk in, 2786, 2798-2800, 2803-05, 2805 noise and, 2789-90, 2790 optical fiber and, 2782-84 optical network unit and, 2780-81 profiling in, 2801-2802, 2802 quadrature amplitude modulation and, 2791, 2801 Reed-Solomon coding and, 2800 reference configurations in, 2785-86, 2787 robustness of, 2787-90, 2789 simulations of, 2788-89 spectral plans for, 2787, 2788 spectrum allocation in, 2786-90 splitters for POTS and, 2785-86 standards for, 2779-84, 2791 timing advance in, 2796-98, 2797 unbundling issues and, 2785 windowing of extra suffix and prefix in, 2799-2800, 2799 very long baseline interferometry, 1927 very low frequency, 208 very small aperture terminal, 879-880, 1247, 2656 vestigial sideband, 1478, 1826 vestigial sideband AM, 133, 136-137, 137, 140 video asymmetric DSL and multimedia transmission in, 1576 bandwidth reduction (see also bandwidth reduction techniques for video service), 232–237 batching in, 234–235, **234** coding for (see also image and video coding), 1025-62 H.324 standard for, 918-929, 919 image sampling and reconstruction and, 1079-94, 1081-92 microelectromechanical systems and, 1349 multimedia over digital subscriber line and, 1576 orthogonal frequency division multiplexing and, 1867, 1878 patching in, 233-234, 234 periodic broadcasting in, 235-236, 236 piggybacking in, 232-233, 232 statistical multiplexing and, 2424-32, 2425 streaming, 2432-41 traffic modeling and, 1672, 1672 unequal error protection coding and, 2764-66, 2765 wireless MPEG 4 videocommunications and, 2972-81 video coding, terrestrial digital TV, 2552 video compression, community antenna TV, 522, 525 video on demand, 232-237, 1558 video signals, in underwater acoustic communications, 37 video streaming, differentiated services, 675 videoconferencing mobility portals and, 2195 session initiation protocol and, 2202 speech coding/synthesis and, 2340, 2827 virtual channel connections, 1658 virtual channel identifier, 201, 206, 270, 549, 550 virtual channels, ATM, 200, 205, 550 virtual circuit deflection protocol, 553 virtual circuit emulation, multiprotocol label switching, 271 virtual circuit identifier, ATM, 264 virtual circuit switching, shallow water acoustic networks, 2208 virtual circuits, 116, 207, 264, 270, 550, 1635, 2211 virtual clock, flow control, traffic management, 1660 virtual containers, synchronous digital hierarchy, 2497 virtual home environment, IMT2000, 1101 virtual LAN, 1282, 1284, 1721-22 virtual path identifer, 200, 201, 206, 264, 270, 549, 550 virtual paths, 200, 205, 264, 273, 550 virtual point to point bit pipe, 539 virtual private networks, 2807-15, 2807 access, 2707-08 ATM and, 273 authentication header in, 2811, 2811

authentication in, 2810 border gateway protocol and, 2809 broadband and, 2663–64, 2664 customer edge in, 1599 eavesdropping and, 2810 encapsulating security payload in, 2810-11, 2810, 2811 encapsulation in, 2708-09 firewalls and, 2809 forwarding in, 2808 generic routing encapsulation and, 2808 hijacking and, 2810 integrated services digital network and, 2808 Internet key exchange protocol and, 2812-14, 2813 Internet protocol and, 2809-14 Internet Security Association and Key Management Protocol and, 2813-14 Internet service providers and, 2808 intranets and extranets in, 1163-72, 1165, 2807 IP in IP encapsulation in, 2808 IP networks and, 2808 IPSec and, 2810-14, 2810 layer 2 forwarding in, 2808 layer 2 tunneling protocol in, 2809 link layer (layer 2), 2808 man in the middle attacks and, 2810 multimedia networks and, 1568 multiprotocol label switching and, 1591, 1599-1600, 1600, 2809, 2809 network layer (layer 3), 2708-09 point to point tunneling protocol in, 2808 provider edge in, 1599 reservation protocols and, 2809 routing and, 2809 security and, 1165-69, 2809-14 spoofing and, 2809-10 subnet to subnet, 2707-08 traffic engineering and, 2809 transport mode in, 2811-12, 2812 tunneling in, 2707-08, 2809, 2811-12, 2812 virtual topologies, routing and wavelength assignment in WDM, 2100-01 virtual tributaries, SONET, 2488-89, 2491 virtual wire service, differentiated services, 674 visible region, antenna arrays, 144 visitor location register, 906, 2987 visual texture coding, image and video coding, 1050 Viterbi algorithm/decoder in acoustic modems for underwater communications, 19 adaptive equalizers and, 81, 90 bit interleaved coded modulation and, 280 compression and, 644 continuous phase frequency shift keying and, 597 continuous phase modulation and, 591 convolutional coding and, 598, 599-600, 600-602, 601 direct sequence CDMA and, 1196-1210 finite traceback Viterbi decoding in, 602, 602 hidden Markov models and, 961 high rate punctured convolutional coding and, 979-993 iterative decoding and, 560 magnetic recording systems and, 2259, 2260, 2265 magnetic storage and, 1330-1331, 1332-1333 multiple input/multiple output systems and, 1455 partial response signals and, 1932, 1933 path metrics of, 600 sequential decoding of convolutional coding and, 2140, 2156, 2161-62 serially concatenated coding and, 2164 simulation and, 2287 soft output Viterbi algorithm and, 2295, 2297-99, 2298, 2302 space-time coding and, 2326 tailbiting convolutional coding and, 2515 trellis coded modulation and, 2627 trellis coding and, 2647, 2648 turbo coding and, 2705, 2714 turbo trellis coded modulation and, 2738, 2743 unequal error protection coding and, 2766-67

1619 Viterbi pruning, in channel modeling, estimation, tracking, 418 VMTP, 2616 VOACAP software, radiowave propagation, 2066 vocoders (see also speech coding/synthesis), 2350-51, 2350, 2819-29 adaptive multirate coder in, 2828 code excited linear prediction and, 2820-29, EFR algorithm in, 2827 enhanced variable rate coder and, 2827 full rate and half rate standards for (Japanese), 2827 global system for mobile and, 2819-20, 2827 linear prediction in, 2820-23, 2821, 2822 long term prediction in, 2823-24, 2825 mean opinion score and, 2819-20 MPEG compression and, 2819 regular pulse excitation algorithm in, 2824 relaxed CELP and, 2827 selectable mode vocoder and, 2827-28 skyphone and, 2824 source system models for, 2820-22 spectral representation of signal for, 2820 standards for, 2819 third-generation CELP and, 2827-28 variable rate CELP and, for CDMA, 2826 videoconferencing and, 2827 voice over IP and, 2827 voice communications community antenna TV and, 512, 523-524 satellite communications and, 880 traffic modeling and, 1671 voice activity detection, 909, 2355 voice coding, IS95 cellular telephone standard, 350, 354 voice over IP (VoIP), 274, 544 satellite communications and, 2121 session initiation protocol and, 2197, 2198 speech coding/synthesis and, 2340, 2354, 2355, 2827 voice recognition (see automatic speech recognition) voice switching, 1906 VoiceXML mobility portals, 2192 voicing, 2360 Volta, 1477 voltage controlled clock, 2029 voltage controlled oscillator frequency synthesizers and, 830, 836, 843, 860 pulse amplitude modulation and, 2027 quadrature amplitude modulation and, 2056 voltage reflection coefficient, active antennas, 56 voltage standing wave ratio, microstrip, 1360, 1363, 1366, 1367, 1390 voltage waves, active antennas, 56 voltages, powerline communications, 1995, 1996 Volterra-based predistorer, 533, 534-535, 534 vortex induced vibration, 20 Vxx standards for modems, 1498-99 Wagner coding, BCH coding, binary, and, 252 Walker star or polar constellations, 1250, 1250 Walsh coding, mobile radio communications and, 1482 Walsh functions cdma2000 and, 362 IS95 cellular telephone standard and, 350, 354, 356-357 Walsh sequences, polyphase sequences and, 1976 Walsh-Hadamard coding, sequences, transforms Golay complementary sequences and, 897-898 multicarrier CDMA and, 1526 signature sequence for CDMA and, 2282-83 vector quantization and, 2126 Wang skywave method and, 2062-64 WAP Forum, 2193, 2899 Ward's method in quantization, 2129-30 water filling in time, wireless multiuser communications systems and, 1606 water pouring result, rate distortion theory and, 2069 water vapor and absorption, millimeter wave propagation and, 1437, 1437 waterfall region, serially concatenated coding and, 2166-67, 2167

watermarking, 1077 waveform analysis, 2575-79, 2576 waveform coding, 2830-37, 2830 adaptive transform coding and, 2837 analog to digital conversion and, 2830 band limited waveforms and, 2831 channel optimized coding and, 2830 codebooks for, 2834 compander and, **2834**, 2834 delta modulation and, 2835 differential pulse code modulation in, 2835, 2835 digitization and, 2830-34 discrete cosine transform and, 2837 discrete Fourier transform and, 2837 discrete Hadamard transform and, 2837 discrete Walsh transform and, 2837 encoding in, 2834-37 Hotelling transform in, 2837 Karhunen-Loeve transforms in, 2837 Nyquist rate and, 2831 pulse code modulation and, 2834-35, 2834 quantization and, 2830, 2832-34 reconstruction in, 2832 sampling and, 2830-32, 2831, 2837 scalar quantization in, 2833 signal to quantization noise in, 2830, 2833 spectral coding and, 2835–37 speech coding/synthesis and, 2341-43, 2819 subband coding and, 2836, 2836 threshold sampling in, 2837 transform coding, 2836-37, 2837 vector quantization in, 2833-34 zonal sampling in, 2837 waveform interpolation coding, speech synthesis/coding and, 1301-02 waveform level simulation, 2285, 2285, 2285 waveguide dispersion, optical fiber and, 1711 waveguide grating filters, 1723, 1723, 1727-28, 1727, 1728 waveguide grating router, 1786 waveguides (microwave waveguides), 1390-1423 active phase aray antenna and, 1391, 1391 airgap matching networks in, 1410-11, 1410, 1412 AMOS 8 feed for, 1392, 1392 antennas as, 169, 179, 180, 184, 187 aperture admittance in, 1406-07, 1406 aperture of, 1406-09, 1406, 1408, 1409, 1419-20 applications for, 1390-1393 attenuation and dielectric losses in, 1405, 1405 bandwidth of, 1390 beamforming network and, 1393 calibration of, 1415-16, 1415 carrier to interference ratio (CIR) in, 1416 cutoff frequency in, 1395, 1396 DFH-3 feed for, 1392, 1392 dielectric filled, 1401-05, 1401, 1403, 1404, 1411-16, 1413-16 discontinuities in. 1412-13 dual polarized, 1416-17, 1417 E plane stepped DFW, 1411-16, 1413-16 Earth Observation Satellite use of, 1391, 1392 eigenfunction and eigenvalues in, 1395 electric fields in, 1394 electromagnetic interference and, 1390 electromagnetic wave mode propagation and, 1390 evanescent waves in, 1395 excitations of modes in, 1397-1405, 1398 external fieds in, 1407-09, 1408, 1409 extremely low frequency in, 758-780 feeds for, 1392–1393, 1392 filtering in, highpass, 1390, 1416–17, **1417** finline transition in, 1399–1400, **1400** frequencies in, 1390, 1395, 1396 fundamental or dominant mode in, 1390 grazing angles in, 1416 Helmholtz (scalar wave) equation and, 1394 horn antennas and, 1006–17, **1006**, 1392, **1392** impedance in, 1395, **1395**, 1398–99, **1399**, 1401–03, 1403 impedance matching in, 1398-99, 1399 IntelSAT, 1392, 1392

internal fields in, 1406 leaky wave antennas and, 1235-36, 1241 lumped circuit network in, 1413 magnetic fields in, 1394 matching network technique in, 1409-11, 1409, 1410 method of moments in, 1420 microstrip end launcher in, 1400, 1400, 1404, 1405, 1401-05, 1401, 1402 microstrip E-plan probe and, 1399, 1399 microstrip/microstrip patch antennas andarrays of, 1373-1374 millimeter wave antennas and, 1426-28, 1427, 1428, 1434 miniaturization technique for, 1405-11, 1406 Nahuel horn antennas and, 1392, 1392 nonradiative dielectric type, 1390 optical couplers and, 1699 optical crossconnects/switches and, 1704, 1704 optical modulators and, 1742 photodetectors and, 1003-04 polarization in, 1390, 1416-17, 1417 power in, 1396-97 radiating slot transition in, 1400, 1400 radiation patterns in, 1417-21, 1418-22 rectangular type, 1393-1405, 1393 reflection coefficient in, 1401, 1402 ridged waveguide transition in, 1400, 1400 satellite and, 1391-1392, 1392 shape and configuration of, 1390-1393 signal to noise ratio and, 1416 slotted, in array, 1391, 1391, 1392 space division multiple access and, 1416 space factor and, 1419 standing waves in, 1394 step ratios in, 1414 surface acoustic wave filters and, 2446-47, 2447 synthetic aperture radar and, 1393, 1393 tapered microstrip transition in, 1400-01, 1400 transverse electric modes in, 1393-96 transverse magnetic modes in, 1393 traveling waves in, 1394 TV SAT horn feed for, 1392, 1392 voltage standing wave ratio in, 1390 wavelength in, 1396, 1396 wide angle impedance matching sheet in, 1391 wavelength active antennas and, 49, 49 millimeter wave antennas and, 1423 optical fiber and, 434, 1714 photodetectors and, 994, 994 waveguides and, 1396, 1396 wavelength assignment, wavelength division multiplexing and, 2097-2105 wavelength converters optical cross connects/switches and, 1799 optical signal regeneration and, 1760 routing and wavelength assignment in WDM and, full, limited, fixed, 2099-2100, 2099 wavelength division multiplexing and, 756, 2840 wavelength dependent couplers, 1761 wavelength division multiple access medium access control and, 1558 optical fiber and, 1808 wavelength division multiplexing, 1824, 2097-2100, 2098, 2653, 2838-46, 2839 all optical network and, 2843-45, 2843 ALOHA protocols and, 2842 asynchronous transfer mode and, 2845, 2864 attenuation in, 2869 bandwidth in, 2865 bidirectional self-healing ring in, 750, 751 BISDN and, 273 bit error rate in, 655 bottlenecks in, 654-655 channels in, 650 chromatic dispersion and, 2869 client and server in, 650-657, 650-656 coarse, 2862 cross gain modulation in, 756 cross phase modulation in, 756 dense WDM and, 748-757, 2271-73, 2862

dispersion in, 2869 dynamic provisioning in, 2865 elements of, 2839-41 enterprise system connectivity in, 2865 erbium doped fiber amplifiers and, 2839, 2869 Ethernet and, 1507 failure and fault detection/recovery in, 1633-34 fiber distributed data interface and, 2864 filtering in, 2869 fixed tuned devices in, 2840-41 four wave mixing in, 756 frequency division multiple access and, 829 frequency division multiplexing and, 2838 frequency selective switches in, 2840, 2841, 2840 future of, 2845 Gigabit Ethernet and, 2864 Internet protocol and, 2864 internetworking in, 654 IP networks and, 2845 joint optical and electronic networks in, 2845 lasers and, 1779 local area networks and, 2841-42, 2842 medium access control and, 1551-52, 1551, 2842 metropolitan area networks and, 2862-73, 2863 multicasting in, 655 multihop mode in, 1551 optical add drop multiplexer in, 651 optical add drop multiplexers and, 651, 748-749, 751-756, 752, 754, 755, 2839-40, 2840, 2864, **2864**, 2867–71, **2868**, **2871** optical communications systems and, 1484–85, **1485**, 1490–91, **1490**, 1709, 1714, 1719–21, **1720**, 1768, 1769, 1841, 1841, 2614, 2615, 2838-46 optical couplers and, 1699 optical cross connects/switches and, 1701, 1797–1808, 2867–71, **2868**, 2864, **2864** optical filters and, 1723, 1731-32 optical modulators and, 1741 optical multiplexing and demultiplexing and, 1748 optical signal regeneration and, 1759 optical sources and, 1775 optoelectronic regenerators in, 2863 passive star topology in, 651, 651 performance of, 654-656 photonic analog to digital conversion and, 1968 polarization and, 2869 receiver for, 651 ring topologies and, 1636, 1636 routing and wavelength assignment in, 2097-2105, 2845 scalability in, 2865 scheduling in, 656-657 self-healing rings in, 750, 751 semiconductor optical amplifiers and, 756, 2869 signal to noise ratio, 2869 signaling in, 653-654 single hop mode in, 1551 solitons and, 1768, 1769 SONET and, 2838-46, 2838, 2839, 2863-73 space division multiplexing and, 2863 synchronous digital hierarchy and, 2863-73 time division multiplexing and, 2864 topologies for, 651 transmitter for, 651 transparency in, 2864 turbo product coding and, 2727 unicasting, 654-655 wavelength assignment in, 2844-45 wavelength converters and, 756, 2840 wavelength routers in, 2839-40, 2840, 2864 wide area networks and, 2842-45, 2843 wavelength of sound, 30, 30 wavelength routers/routers, 1798 optical cross connects/switches and, 1799-1800 optical multiplexing and demultiplexing and, 1749 wavelength division multiplexing and, 2839-40, 2840, 2864 wavelength selective optical crossconnects, 1703, 1703 wavelength-selectable lasers, 1779-80, 1780 wavelet coding, image compression and, 106970

wavelets, wavelet transforms, 2846-62

Web crawlers, paging and registration in, 1915 websites, cost of, 1169-71 Weibull fracture probability distribution, 438 weighted fair queue, 1564, 1565, 1660 weighted round robin, 1564, 1565 weighted TDM, 1552 weighting antenna modeling and, 175, 176 automatic repeat request and, Hamming weight, weight enumerator) in 225-226 synaptic weight, neural networks and, 1676 Welch bound Gold sequences and, 901 ternary sequences and, 2543, 2544, 2545 West satellite network, 2112 White Book, 1736 whitened matched filter, adaptive equalizers and, 89 Wichmann-Hill algorithm, random number generation and, 2292 wide angle impedance matching sheet, waveguides and, 1391 wide area networks, 2461 BISDN and, 271-272 broadband and, 2663-64 Ethernet and, 1512 fault tolerance and, 1632 Gigabit Ethernet and, 1509 indoor propagation models for, 2012-21 intranets and extranets in, 1163-72, 1165 IP telephony and, 1172-82, 1173 multicasting and, 1532 optical fiber systems and, 1714, 1840 powerline communications and, 1997-98 random number generation and, 2269 reliability and, 1632 routing and wavelength assignment in WDM and, 2098 wavelength division multiplexing and, 2842-45, 2843 wireless infrared communications and, 2925 wireless multiuser communications systems and, 1602 wide sense stationary, 393, 394, 1667 wideband communications in channel modeling, estimation, tracking, 409, 411 indoor propagation models and, 2017–18 speech coding/synthesis and, 2341 ultrawideband radio and, 2754-62 wideband CDMA, 1196, 2391, 2873-83, 2954 admission control and, 126 cell planning in wireless networks and, 372, 386 chann/in channel modeling, estimation, tracking, 409 diversity and, 733-734 filtering in, 2878 frame structures in, 2877, 2877 IMT2000 and, 1096, 1104-05, 2873-74 mobile radio communications and, 1483, 1483 modulation in, 2878 multidimensional coding and, 1548 physical channels in, 2876-79 physical random access channels in, 2879, 2879 power control and, 1986 quadrature phase shift keying and, 2878 random access in, 2881–82 satellite communications and, 2116 scrambling codes in, 2878 signature sequence for CDMA and, 2282 spread spectrum and, 2400 spreading in, 2877-78 synchronization channels in, 2878-79 transport channels in, 2879-82 universal mobile telecommunications system and, 2873-74 UTRAN and, 2873-74 wireless multiuser communications systems and, 1602, 1608 Wiener filters (see also minimum mean square error), 686, 701-702 acoustic echo cancellation and, 6 in channel modeling, estimation, tracking, 412, 413 packet rate adaptive mobile receivers and, 1888, 1892 Wi-Fi, 1288, 2400

and, 1373, 1373, 1382 window regions millimeter wave propagation and, 1433, 1434 optical fiber and, 1708 window size constrained coding techniques for data storage and, 575 flow control and, 545, 1627 orthogonal frequency division multiplexing and, 1872–73, **1873** windowed Fourier transform, 2848-49 Window-Hoff rule, neural networks and, 1677-78 wire modeling for antennas and, 174-175, 174, 175, 182–183, **183** wired equivalent privacy, 1155-56, 1286, 1287-88 wireless application environment, 2901-02 wireless application protocol, 1156, 2899-2906 Apache servers and, 2900 bearers on air interface and, 2903-04 common gateway interface and, 2900 extensible HTML and, 2899 extensible markup language and, 2899 general packet radio service and, 866 global system for mobile and, 908 handheld device markup language and, 2899 hypertext markup language and, 2900 hypertext transfer protocol and, 2899 IMT2000 and, 1100 keyboards for, 2900-01 man machine interface and, 2901 mobility portals and, 2190, 2191, 2192-93, 2193, 2194-95 personal digital assistants and, 2899 protocol stack for, 2901-04, 2902 security and, 2194-95 smart messaging and, 2899 tools and applications for, 2904-05 uniform resource locators and, 2901 WAP Forum and, 2899 wireless application environment and, 2901-02 wireless datagram protocol, 2902-03 wireless markup language and, 2899 wireless session protocol and, 2902 wireless telephony application and, 2901 wireless transaction protocol and, 2902 wireless transport layer security and, 2902 wireless datagram protocol, 2902-03 Wireless Ethernet Compatibility Alliance and, 1288 wireless identity module, 2195 wireless infrared communications, 2925-31, 2925 applications for, 2925 bandwidth efficiency and power efficiency in, 2927 bit error rate in, 2927 building to building systems for, 2929-30 challenges for, 2930 channel impulse response in, 2928-29 coherent detection in. 2926 direct detection in, 2926 error control coding in, 2927-28 future of, 2930 IrDA and, 2929 modulation/demodulation in, 2927-27 on off keying and, 2927, 2928 radio vs., 2930 receivers and transmitters for, 2926 return to zero in, 2927, 2928 safety and, 2927 signal to noise ratio, 2927 standards and systems for, 2929-30 wavelength and noise in, 2926-27 wireless LANs and, 2929 wireless IP suite enhancer, 1233 wireless IP telephony (see also IP telephony), 2931-41 wireless LAN, 1284-89, 1285, 2391, 2678-82, 2680, 2941-47 access points (AP) in, 1285 ad hoc networks in, 1285 antennas, 190 association, disassociation, reassociation in, 1287 asynchronous transfer mode and, 2681

Wilkinson dividers, microstrip/microstrip patch antennas

attacks on, 1288 authentication in, 1287, 1288 bandwidth and, 2678 Barker coding in, 2942 basic service set in, 1285 Bluetooth and, 308, 1289 BodyLAN and, 2681 carrier sense multiple access and, 346, 1286, 2678, 2682, 2945 cdma and, 2679 channelization in, 2944-45 complementary key coding in, 2943 coverage and, 2678 digital audio/video broadcasting and, 2941 digital enhanced cordless telephony and, 1289 direct sequence spread spectrum and, 1285, 2678, 2842-43, 2842 distribution system for, 1285 extended service set in, 1285, 1286 fragmentation in, 1287 frequencies for, 2678-79 frequency hopping spread spectrum in, 1285 future of, 1289 hidden node problem in, 1286-87, 1286 HiperLAN and, 2681, 2682, 2941, 2945 HomeRF and, 1289 IEEE 802.11g high speed, 1289 independent BSS in, 1285 interference and, 2678 local multipoint distribution services and, 1269 MAC data frame formats in, 1286, 1286 marketing strategies for, 2679-80 media access control and, 5, 1285-87, 1343, 2942 military use of, 2680-82 multimedia mobile access communication and, 2941 network allocation vector in, 1287 orthogonal frequency division multiplexing and, 1288-89, 1867, 2941, 2942, 2943-45, 2944, 2945, 2946 packet binary convolutional coding and, 2946 physical layer in, 1285 power management in (awake, doze state), 1287 roaming in, 1287 satellite communications and, 2118 security and, 1155–56, 1287–88 service providers and, 2680-82, 2682 software radio and, 2314 spread spectrum in, 1285, 2399 standards for, 2682, 2945-46 topologies for, 1285, 1285 trends in, 2677-92 unlicensed bands for, 2678-79 unlicensed national information infrastructure and, 2941 Wi-Fi and, 1288 wired equivalent privacy and, 1286, 1287-88 Wireless Ethernet Compatibility Alliance and, 1288 wireless infrared communications and, 2929 wireless multiuser communications systems and, 1602 wireless packet data and, 2982 wireless local loop standards and systems, 2947-59, 2948, 2949 wireless markup language, 2899 wireless MPEG 4 videocommunications, 2972-81 wireless packet data, 2981-90, 2983 architectures for, 2982-84 coverage and mobility support in, 2985 data rates in, 2982 delay in energy efficiency in, 2985-86, 2985 frequencies for, 2982 gateway GPRS support node in, 2983-84, 2983, 2988 general packet radio services and, 2983-84, 2983, 2988 header compression in, 2987 home location register in, 2987 interference in, 2982 internet control message protocol and, 2988 IP addressing and, 2988 link adaptation and incremental redundancy in, 2986

wireless packet data (continued) local area networks and, 2982 logical link control and, 2983 losses in, 2985 medium access control in, 2982 mobile IP and, 2988-89, 2989 mobility management in, 2987-89, 2987 peak picking scheduling in, 2986-87, 2986 quality of service, 2984 radio link control in, 2982, 2984 security in, 2985 serving GPRS support node in, 2983-84, 2983, 2988 throughput in, 2984-85 transmission control protocol and, 2984, 2989 visitor location register in, 2987 wireless LANs and, 2982 wireless PAN (see also personal area networks), 502, 508, 2682 wireless routing protocol, ad hoc wireless networks and, 2886, 2888 wireless sensor networks, 2990-96 wireless session protocol, 2902 wireless systems (see also Bluetooth; land mobile satellite communications; mobile radio communications), 208, 308, 408, 1223-34 ad hoc networks in, 2883-99 adaptive receivers for spread-spectrum system and, 96-97 admission control in (see admission control in wireless networks) analysis techniques for, 2919-24 angle of arrival in, 2689-90, 2689 antennas for (see also antennas for mobile communications), 169, 179-180, 184, 188-200 asynchronous transfer mode and, 2906-15, 2907 bit error rate, 2923, 2924 bit interleaved coded modulation and, 276 blind equalizers and, 296-297 Bluetooth and, 307-317 broadband access (see broadband wireless access) burst switched networks and, 122 carrier sense multiple access and, 346 carrier to interference ratio in, 121, 379, 379 cdma2000 and, 385-386 cell planning in, 369-393 cells in, 369-393 in channel modeling, estimation, tracking, 398-408 channel assignment problem and, 382-383, 383 channel borrowing and, 125, 125 channel tracking in (see channel tracking in wireless systems) circuit switched networks and, 122, 371 coding division multiple access and, 372 common packet channel switching and, 123 compression of data and, 371 coverage area and, 372 data services on, 371 design of, 2915-25 diversity reception in, 2919-20 Doppler effect, Doppler spreading in, 2916-18., 2917, 2918 equal gain combining in, 2920, 2921-22 Erlang B blocking in, 379-380, 379, 380 extremely low frequency in, 758-780 fading in, 2915, 2916-18 first-generation sysetms in, 370 fourth generation systems in, 371-372, 391-392 free space optics in (see free space optics)

frequencies for, 208 frequency assignment problem and, 382-383 frequency division duplex and, 385-386 geolocation of, indoor, 2688–90, **2689** global system for mobile and, 369-372, 377-383 global vs. local parameters for, 372 grade of service and, 379-380 guard channels and, 124, 124 H.324 standard for, 918-929, 919 history of, 369-372 home area network and, 2685-88, 2687 home networking and, 2684-88, 2685 IMT2000 and, 386, 392, 1094-1108 indoor propagation models for, 2012-21 interference and, 121, 377-380, 1151-30 Internet and, 371 IP networks and, 392 local multipoint distribution services in, 1268-79 location in, 2959-72 lossy compression and, 371 low density parity check coding and, 1316 maximal ratio combining in, 2920 media access control and, 1342-1349 microelectromechanical systems for, 1349-1356 mobility portals and services for, 2190-96 modulation and, 371 Monte Carlo simulation and, 2916 multibeam phased arrays and, 1513-21, 14 multimedia service in, 371 multipath and, 2916-18 multiple antenna transceivers for (see also multiple antenna transceivers), 1579-90, 1580 multiple input/multiple output systems and, 1450-56, 1450 multiuser systems (see multiuser wireless communication systems) noise and, 2915 nonlinear least square in, 2690 optimization in, 372 orthogonal frequency division multiplexing and, 2916 outage probability in, 2922 packet switched networks and, 122, 371 paging and registration in, 1914-28 parabolic and reflector antennas and, 2080 planning for, 372 power control and, 121-122 quality of service and, 372, 379, 379, 1450, 2915 queuing priority and, 124-125, 125 radio resource management and, 2088-97, 2089 received signal phase in, 2690 received signal strength in, 2690 satellite communications and, 2111-22 second generation systems in, 370-371, 377-383 security and, 1155-56, 1646 selection combining in, 2920-21 shadowing and, 2922 Shannon or channel limit in, 2915 signal correlation in. 2917-18 signal to interference ratio in, 121 signal to noise ratio, 121-122, 2915, 2919, 2921 space-time coding for, 2324-32, 2324 spatiotemporal signal processing in, 2333-40, 2333 standards for, 371 system model for, 2922-23, 2923 TDSCDMA and, 385-386 third-generation, 125-126 third-generation systems in, 371, 385-391 time difference of arrival in, 2690

time division duplex and, 385-386 time division multiple access and, 377, 380 time of arrival in, 2690 traffic computation for, 380 trellis coded modulation and, 2922 trellis coded PSK, 2922 Universal Mobile Telecommunications Systems and, 371-372, 385-391 voice services on, 371 wideband CDMA and, 372, 386 wireless application protocol and, 2899-2906 WRC2000 and, 392 wireless telephony application, 2901 wireless transaction protocol, 2902 wireless transmission control protocol, 1233 wireless transport layer security, 1156, 2902 wiretapping, 1646 Woods Hole Oceanographic Institute, acoustic telemetry in, 24 Woodward-Lawson method, 158-159, 159, 187 World Administrative Radio Conference, 1478 world wide web (see also Internet), 271-274, 540 World Wide Web Consortium, 2193, 2899 worst case fair weighted fair queuing, 1660 WRC2000, 392 write once read many devices, 1319, 1737 write process CDROM and, 1734 digital magnetic recording channel and, 1323-1324 digital versatile disc and, 1738 hard disk drives and, 1320 magnetic recording systems and, 2249 magnetic storage and, 1320, 1326 optical memories and, 1733 X.25, 546 X.509, 1649 X2 protocol, 1498, 1499 Xerox PARC and local area networks and, 1279 Xon/Xoff modems and, 1496, 1497 XORing cryptography and, 608-609 Xpress transport protocol and, 2616, 2617, 2620 Yagi-Uda, antenna arrays and, 161, 161, 169, 187 Ymodem, 1497 Z transforms, digital filters and, 690-691 zenith angle, 2067 zero crossing rate, speech coding/synthesis and, 2371 zero force projection, wireless transceivers, multiantenna and, 1588 zero forcing equalizers and, 81, 82-83, 88 tapped delay line equalizers and, 1690 multiple input/multiple output systems and, 1455 zero input response, 2347 zero knowledge, authentication and, 614 zero state response, 2347 zero tree coding, 1046-47 zigzag scanning, image and video coding and, 1046 Zmodem, 1497 zonal sampling in waveform coding and, 2837 zone-based hierarchical link state, 2890 zone bit recording, 1321 zone of authority, 548 zone routing protocol, 2889 zoned linear velocity, 1735