FREQUENCY MODULATION

Frequency-modulated waveforms are commonly utilized for information transmission in radio communications, as well as for environmental sensing in radar, sonar, and bioengineering. In many of these situations, the desired information is extracted from the received signal by monitoring one or more dominant frequencies in the signal and examining their variation as a function of time. The process by which a local approximation to a signal's frequency is obtained is known as instantaneous frequency estimation (IFE).

In this article, concepts relating to instantaneous frequency, along with algorithms for its computation, are reviewed. First, several means by which instantaneous frequency is commonly defined are discussed, and the relationships between instantaneous frequency and time–frequency distributions are explored. Next, several measures of performance commonly used to evaluate instantaneous frequency algorithms, such as the Cramer–Rao lower bound, are examined. Finally, a number of algorithms which have recently been suggested for IFE are summarized. Despite the relative maturity of frequency modulation in the field of radio communications, the field of IFE is a growing one, and one in which research is still quite active.

INSTANTANEOUS FREQUENCY ESTIMATION: BACKGROUND AND DEFINITIONS

The instantaneous frequency of a signal can be defined in several different ways. Two of the most popular definitions relate the instantaneous frequency to time–frequency distributions and to the analytic signal. Either of these definitions can satisfy intuition in certain situations, but yield puzzling results in others. Nevertheless, useful estimates of the instantaneous frequency can usually be obtained for narrowband signals, and to a lesser degree for certain wideband signals. In this section, relationships between instantaneous frequency, time–frequency distributions, and the analytic signal are discussed.

Time–Frequency Distributions

The concept of frequency has long played a major role in the analysis of signals. Through the Fourier transform, a signal may be decomposed into a continuum of complex exponentials. In fact, the basis functions of the Fourier transform are pure tones of *infinite* time extent. However, when the spectral composition of a signal varies as a function of time, the Fourier transform no longer provides a simple spectral description of the signal. Instead, a *time–frequency distribution* yields more insight into the signal's behavior. The most common example of time–frequency analysis—the printed musical score—has existed for hundreds of years. With a musical score, it is possible to denote the tones that are present in an arrangement at discrete intervals in time. In the following paragraphs, a short discussion of the key developments in time–frequency analysis that have been obtained in the last 50 years is provided. Among the topics explored are: Gabor's time–frequency distribution, the short-time Fourier transform, perfect reconstruction filter banks, the wavelet transform, and Cohen's class of time–frequency distributions. The relationship between instantaneous frequency and time–frequency distributions is then discussed.

The development of the first algorithm for time–frequency analysis of an arbitrary signal is generally credited to Gabor (1). His work was motivated by a desire to define the information content of signals. He considered the time–frequency representation of a signal as a "diagram of information," with areas in the two-dimensional representation being proportional to the amount of data that they could convey. Gabor suggested that the time and frequency characteristics of a signal $x(t)$ be simultaneously observed with the expansion

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} g_{mn} \psi_{mn}(t)$$

where $\psi_{mn}(t)$ is expressed in terms of an elementary signal $\psi(t)$ with

$$\psi_{mn}(t) = \psi(t - mT)\exp(jn\Omega t)$$

The time and frequency lattice intervals are defined by T and Ω, respectively. Gabor also proposed that the signal with minimum area on the time–frequency plane be used to generate the basis functions for his time–frequency decomposition. Furthermore, he demonstrated that the signal with minimum area, as defined by the product of the signal's root mean square (rms) width in time and frequency, was given by the Gaussian-modulated sinusoid. The concept that the time and frequency widths of a signal cannot be made arbitrarily small simultaneously is a well-known property of Fourier analysis called the uncertainty principle (a term that originated in the physics community). These ideas form the cornerstones of time–frequency analysis.

The growing interest in time–frequency analysis accelerated in the 1970s due to research regarding the short-time Fourier transform (STFT) (2,3). These efforts were motivated by a desire to analyze the time-varying spectral content of speech signals. The STFT is created by inserting a window function $h(n)$ into the computation of the Fourier transform, as expressed by

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x(m)h(n-m)\exp(-j\omega m)$$

The discrete index n varies from $-\infty$ to $+\infty$, while the continuous parameter ω varies from 0 to 2π. For a fixed analysis time n, the window function selects a portion of the original signal for spectral analysis, thereby allowing nonstationary behavior to be observed in a manner impossible with the traditional Fourier transform. At each time instant, the signal segment selected for analysis is formed by the product of the time-shifted window function with the original signal. It is recognized that the result of this operation in the frequency domain is the convolution of the spectral representation of the two functions. Thus, the shape of the window function is fundamental to the STFT results. The rectangular window yields the minimum mainlobe spectral width (and therefore

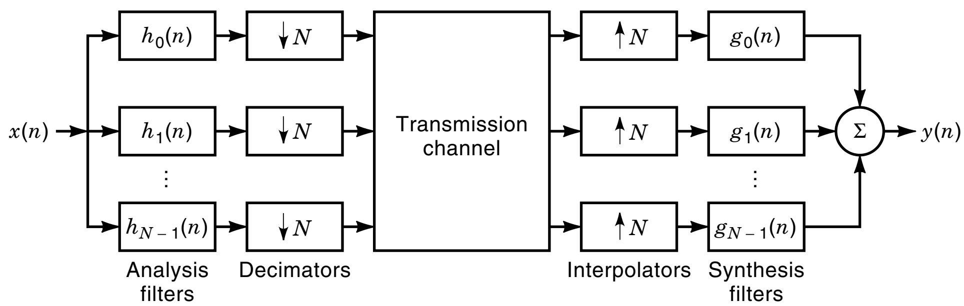

Figure 1. Structure for N-channel filter bank. This tool is frequently employed for time–frequency analysis.

the best frequency resolution) at the cost of large sidelobes (the peak of the largest sidelobe has a magnitude only 13 dB less than the peak of the mainlobe). Other window functions, such as the Hamming window, generate lower sidelobes, at the price of a wider mainlobe. The length of the window function also affects the results produced by the STFT. The longer the time length of the window, the greater the frequency resolution of the time–frequency representation, but the poorer the time resolution.

A second interpretation of the STFT is obtained by examining the structure of the discrete short-time Fourier transform (DSTFT), which is written as

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)h(n-m)\exp\left(\frac{-j2\pi km}{N}\right)$$

The discrete index n varies from $-\infty$ to $+\infty$, while the discrete index k varies from 0 to $N-1$. For a fixed frequency index k, it is seen that the signal of interest is modulated by $\exp(-j2\pi km/N)$ and then convolved with the window function (which typically has a low-pass frequency response). The series combination of the modulator and the window function results in a bandpass filter. The DSTFT can therefore be thought of as being generated by passing a signal through a set of bandpass filters. Furthermore, if the discrete signal is processed with an ideal bandpass filter with a passband equal to π/N, then the output of the filter can be *decimated* by a factor of N (all signal samples are discarded except those corresponding to indices of N, $2N$, $3N$, etc.) with no loss of information. The incorporation of this philosophy with the bandpass filter view of the STFT yields the analysis portion of the *filter bank* structure shown in Fig. 1. The rationale for decimating the output of each of the bandpass filters is to reduce the storage requirements of the filter bank. In a similar fashion, a signal approximately equal to the input signal can be produced with the synthesis structure also shown in Fig. 1. Note that the synthesis network contains *interpolators* (which insert N zeros between each input sample) and window functions known as synthesis filters. In general, the synthesis filters are similar, but not identical, to the analysis filters. In a typical data compression application of a filter bank, the outputs of the analysis section are encoded, transmitted over a channel, and then reconstructed with the synthesis section. Equality between the input and output of the filter bank subject to a finite delay, called *perfect reconstruction,* is achieved only for specific combinations of analysis filters and synthesis filters.

It is evident that filter banks are a natural extension of the STFT. As is the case for the STFT, filter banks may be utilized for instantaneous frequency estimation. Properly implemented, they permit a compact representation of the time-frequency properties of a signal with no loss of information. The characteristics of analysis and synthesis filters that eliminate aliasing, along with the requirements for perfect reconstruction, are thus of interest.

In an actual implementation of a filter bank, the digital analysis filters are of finite length, and thus they cannot form an ideal bandpass filter with unity gain in the passband and zero gain in the stopband. Therefore, some degree of aliasing will occur when the output of each of the analysis filters is decimated as in Fig. 1. If the analysis filters are not correctly designed, the aliased signals propagate through the filter bank and severely limit the quality of the filter bank's output. Croisier et al. (4) examined this problem for the two-channel filter bank shown in Fig. 2, and they derived conditions on the analysis and synthesis filters such that these aliased terms are completely canceled. Filters designed with this approach are called *quadrature mirror filters,* since the analysis filters of a two-channel network are mirrors of one another about $\pi/2$. It is important to note that while the output of a filter bank employing quadrature mirror filters does not contain aliased terms, other magnitude and phase distortions typically exist.

The necessary and sufficient conditions for the design of a perfect reconstruction filter bank were derived by Smith and Barnwell (5). They also proposed an algorithm to construct analysis and synthesis filters which satisfied these conditions, using well-known filter design techniques (6). Although they termed these filters *conjugate quadrature filters,* many researchers consider them to a be a class of quadrature mirror filters. The power of Smith and Barnwell's algorithm is demonstrated by the fact that it is applicable to the two-channel structure shown in Fig. 2, the N-channel structure shown in Fig. 1, and the tree-structured analysis section shown in Fig. 3, as well as to filter banks employing nonuniform decimation and interpolation rates. Additional results regarding the implementation of perfect reconstruction filter banks are included in Refs. 7 and 8.

A closely related topic to time–frequency analysis is *time-scale* analysis, which is provided by the *wavelet transform.* Due to the similarity of the discrete wavelet transform with perfect reconstruction filter banks, as well as the immense number of applications of the wavelet transform that have been investigated over the past 10 years, a brief discussion of its development is included in the following.

In the late 1970s, the French geophysical engineer Morlet derived an alternative to the STFT for time–frequency analysis. The seismic signals of interest to Morlet contained high-

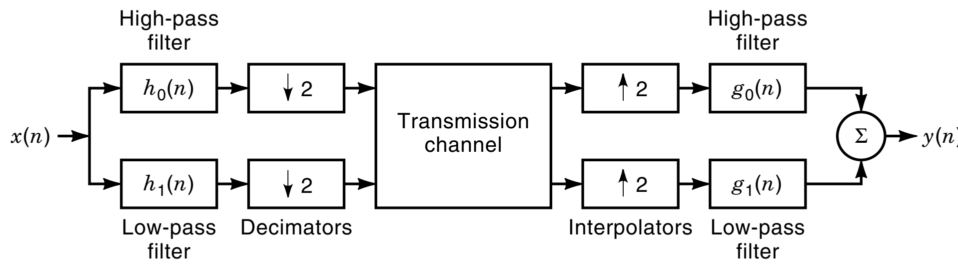

Figure 2. Structure for two-channel filter bank. This simple configuration can be used to construct more complicated structures.

frequency components with shorter timespans than did the low-frequency components. With the STFT, it was impossible to simultaneously obtain good time resolution for the high-frequency components and good frequency resolution for the low-frequency components. Morlet recognized that his goal could be obtained by decomposing the seismic signals not with *translated* and *modulated* versions of an elementary signal (as was done in Gabor's work and the STFT), but with the *translated* and *scaled* versions of an elementary signal. This concept yielded basis functions which contained a constant number of cycles. Morlet chose to call his functions "wavelets of constant shape" (9). Although the term wavelet had been used in the seismic field for a number of years before Morlet's work, it had been used to denote seismic pulses, not a time-frequency tool.

Morlet later collaborated with Grossman to place the wavelet transform on a firm mathematical foundation (10). They defined the continuous wavelet transform as

$$W_x(\tau, a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) m^* \left(\frac{t-\tau}{a} \right) dt$$

where a is a scale factor, and the "mother wavelet" $m(t)$ serves as a window function. The inverse wavelet transform is given by

$$x(t) = c \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{a^2} W_x(\tau, a) m \left(\frac{t-\tau}{a} \right) da\, d\tau$$

Morlet and Grossman used a nonorthogonal basis for the transform consisting of functions very similar to Gabor's Gaussian-modulated sinusoids. An orthogonal basis was developed in 1985 by Meyer.

Structures for the discrete wavelet transform and its inverse were developed by Daubechies (11). The form of the forward structure is shown in Fig. 4 and is seen to be very similar to a pruned version of the tree-structured filter bank shown in Fig. 3. At each level of the discrete wavelet transform, the input signal is passed through a low-pass and high-pass filter. The outputs of the filters are decimated by a factor of two, and the decimated low-pass filter output is again passed to a low-pass and high-pass filter pair. Daubechies derived conditions for the filters such that the structure yields perfect reconstruction, and she used these conditions to generate a set of viable filters frequently referred to as "Daubechies wavelets." Of special interest to Daubechies was the relatively long impulse response of the filter produced by a series of short filters alternated with decimators. She termed the combined impulse response of the low-pass filters alternated with decimators the *scaling function,* and the impulse response of the low-pass filters alternated with decimators and followed by a high-pass filter the *wavelet function.* She showed that as the number of levels in the transform grows large, the scaling function and the wavelet function converge to smooth waveforms, provided that the component filters have sufficient "regularity." In digital signal processing terms, the regularity of a filter corresponds to the number of zeros at z =

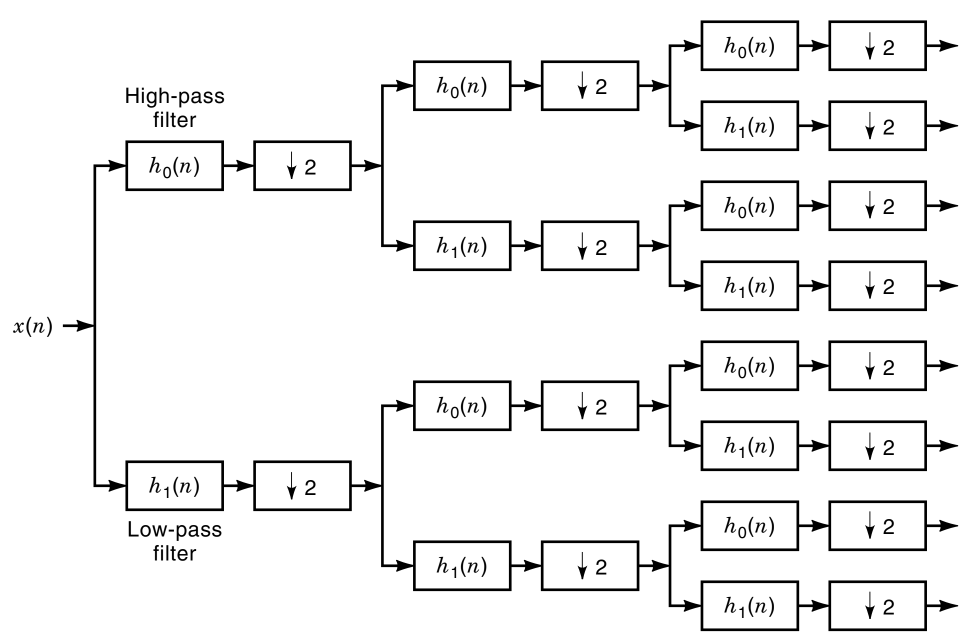

Figure 3. Analysis portion of tree-structured filter bank formed by the sequential application of two-channel filter banks.

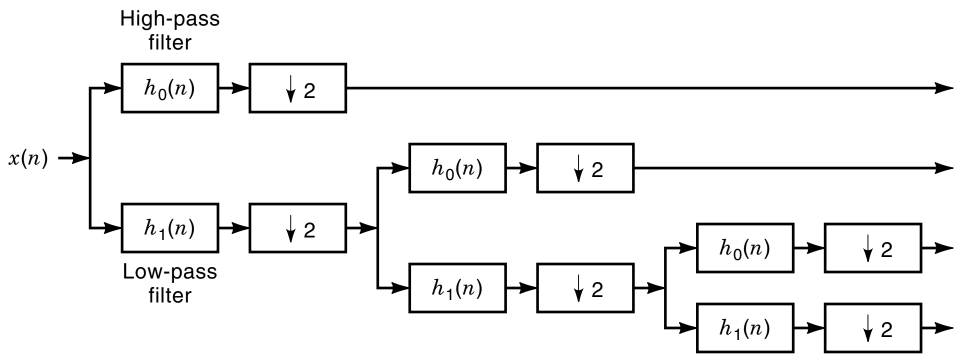

Figure 4. Structure for discrete wavelet transform. Note the similarity of this configuration to the structure shown in Fig. 3.

-1 ($\omega = \pi$) in the filter transfer function. Soon after Daubechies' work was published, Mallat extended her results to two dimensions and applied them to image processing (12). It has since been recognized that Daubechies' conditions for perfect reconstruction are identical to those published by Smith and Barnwell. It has also been recognized that the structure corresponding to the discrete wavelet transform is equivalent to an octave-channel filter bank, a form of which had been investigated earlier for speech-coding (3). Nevertheless, the attention focused on filter banks and time–frequency analysis due to the introduction of the wavelet transform has resulted in an explosion of new developments in these fields that has continued today.

In addition to the time–frequency distributions described above, many others have been developed over the past 50 years. A significant number of continuous time–frequency distributions can be characterized by what is known as *Cohen's class* of distributions (13), which is defined by

$$C_x(t,\omega) = \frac{1}{4\pi^2}\iiint x^*(u-\tfrac{1}{2}\tau)x(u+\tfrac{1}{2}\tau)\phi(\theta,\tau) \exp(-j\theta t - j\tau\omega + j\theta u)\,du\,d\tau\,d\theta$$

where $\phi(\theta, \tau)$ is a two-dimensional function known as the kernel, and $x(t)$ is the signal under consideration. A particular member in Cohen's class is identified by its kernel. For example, the *spectogram,* which is defined as the magnitude squared of the STFT, is a recognized member of Cohen's class of distributions, with a kernel given by

$$\phi(\theta,\tau) = \int h^*(u-\tfrac{1}{2}\tau)\exp(-j\theta u)h(u+\tfrac{1}{2}\tau)\,du$$

where $h(t)$ is the window function defined previously. A discrete form of Cohen's class of time–frequency distributions is examined in Ref. 14.

With respect to a time–frequency distribution, there are two possible means of defining the instantaneous frequency of a signal at a point in time. The instantaneous frequency may be associated with either (1) the peak value of the signal's distribution at that time or (2) the average of the frequencies present in the signal at that time. These approaches are appealing because they permit the introduction of the *instantaneous bandwidth* concept in a natural manner as the spectral spread of energy in the time–frequency plane about the instantaneous frequency. In the following section, it is shown that the instantaneous frequency estimate derived from certain time–frequency distributions is equivalent to the corresponding quantity derived from the analytic signal.

A major benefit obtained by employing a time–frequency distribution for instantaneous frequency estimation is the capability of the distribution to aid in the determination of whether the signal under examination is monocomponent or multicomponent. Monocomponent signals are those which can be shown to possess energy in a contiguous portion of the time-frequency plane. At any point in time, this type of signal exhibits a narrowband characteristic. An example of this type of signal is a sinusoid with a continuous time-varying frequency. Conversely, multicomponent signals are those that can be shown to possess energy in multiple, well-isolated frequency bands at the same instant in time. Speech frequently displays this behavior. It is noted that the above definition of monocomponent excludes signals such as an impulse, which could also be argued to be monocomponent due to its ridge-like time–frequency distribution. Obviously, the identification of a signal as monocomponent or multicomponent is not precise (15); but it is important, as the instantaneous frequency of a multicomponent signal may have no physical meaning (16).

A significant disadvantage of employing time–frequency distributions for IFE is that the construction of the distribution is a computationally complex procedure, even when filter bank structures are utilized. Fortunately, there is an alternative approach for IFE. In many situations, a reasonable estimate of the instantaneous frequency of a signal can instead be obtained from computationally simple operations on its analytic signal. Background for this philosophy is given in the following section. A summary of algorithms which have been suggested for the implementation of this approach are provided in the section entitled [Algorithms for Instantaneous Frequency Estimation.]

Analytic Signals

In this section, the relationships between the instantaneous frequency of a signal and its analytic signal are examined. First, a brief review of the analytic signal is provided. The definition of the instantaneous frequency in terms of the analytic signal is then discussed. Practical issues regarding the computation of the analytic signal are also presented. Finally, the situations for which the estimate of instantaneous frequency obtained via the analytic signal agrees with the estimate obtained from certain time–frequency distributions are examined.

Interestingly, as was the case with time–frequency analysis, the original work done in the area of analytic signals was conducted by Gabor (1). He defined the complex analytic signal $z(t)$ corresponding to a real signal $x(t)$ to be the sum of the signal with a second signal generated via the Hilbert transform

$$\begin{aligned} z(t) &= x(t) + j\boldsymbol{H}\{x(t)\} \\ &= x(t) + jy(t) \end{aligned}$$

The continuous Hilbert transform is defined as

$$\boldsymbol{H}\{x(t)\} = \text{p.v.}\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}\,d\tau$$

where p.v. indicates the Cauchy principal value of the integral. The signal $z(t)$ is similar to $x(t)$, in that for positive frequencies, $Z(f) = 2X(f)$. However, the spectrum of $z(t)$ contains no energy at negative frequencies. Indeed, one technique for deriving the analytic signal of a real signal is to compute its Fourier transform, ignore the spectral components corresponding to negative frequencies, apply the appropriate scaling factor, and then to compute the inverse Fourier transform of the modified signal. Gabor developed the analytic signal concept to aid in the derivation of the signal with minimum time–frequency extent. However, the most extensive application of analytic signals has been in the communications field.

The importance of the analytic signal to the definition of instantaneous frequency can be seen by considering a simple example. Suppose a continuous real signal $x(t)$ is given by

$$x(t) = a(t)\cos(\phi(t))$$

where $a(t)$ represents a time-varying amplitude, and $\phi(t)$ represents a time-varying phase. Since the "frequency" of a sinusoid is defined as the derivative of its phase, the instantaneous frequency of a signal $x(t)$ could be computed with the derivative of $\phi(t)$. This definition appears to agree with intuition. However, when the magnitude of $a(t)$ is bounded by b the signal $x(t)$ can also be expressed as

$$x(t) = b\cos(\tilde{\phi}(t))$$

where $\tilde{\phi}(t) \neq \phi(t)$. Therefore, the postulated definition does not yield a unique instantaneous frequency for the signal $x(t)$. By defining the instantaneous frequency of $x(t)$ to be the derivative of the phase of the corresponding unique analytic signal $z(t)$, this ambiguity can be eliminated (17). Since the analytic signal is complex, it can always be expressed uniquely as

$$z(t) = m(t)\exp(j\theta(t))$$

The instantaneous frequency can thus be uniquely defined as $d\theta(t)/dt$. However, it is not claimed that this definition provides satisfactory results in every scenario.

In practice, discrete sequences $x(n)$ corresponding to samples of the continuous signal $x(t)$ at time instants $t = nT$ are available, and it is desired to form the discrete analytic signal $z(n)$ where

$$z(n) = x(n) + jy(n)$$

The analytic signal $z(n)$ can be obtained in a number of ways, each with its own advantages and disadvantages. The sequence $y(n)$ can be generated by the brute force approach of zeroing the spectral components of $x(n)$ corresponding to negative frequencies. This technique works well for finite data sets, but can be difficult to implement in real-time applications. The sequence $y(n)$ can also be generated by processing the sequence $x(n)$ with a digital filter designed to approximate the Hilbert transform (6). However, the group delay introduced to $y(n)$ by the digital filter must also be introduced to $x(n)$, which can be difficult for noninteger delays. A third approach is to compute the complex sequence $z(n)$ directly, by processing the sequence $x(n)$ with a complex filter constructed by modulating a real low-pass filter by a complex exponential (18).

It is interesting to note the relationship between (1) the instantaneous frequency estimate obtained from a time–frequency distribution and (2) the instantaneous frequency estimate obtained from the derivative of the analytic signal's phase. For continuous signals, it can be shown that the first moment of a time–frequency distribution of Cohen's class is equivalent to the derivative of the analytic signal's phase if the kernel $\phi(\theta, \tau)$ is selected such that

$$\left.\frac{\partial\phi(\theta,\tau)}{\partial\tau}\right|_{\tau=0} = 0$$

Furthermore, for signals with quadratic phase functions, the peak of the time–frequency distribution known as the Wigner–Ville distribution corresponds to the instantaneous frequency (19). Results for the discrete signal case have appeared in Ref. 20 and in Ref. 14.

MEASURES OF PERFORMANCE FOR INSTANTANEOUS FREQUENCY ESTIMATION ALGORITHMS

To evaluate various IFE algorithms, there must be a means of comparing the performance and implementation of a specific algorithm to an alternate approach. In this section, the measures of performance typically used to compare IFE algorithms to one another are discussed. The measures of performance that are considered include both statistical and computational issues. At the conclusion of this section, these criteria are demonstrated by utilizing them to evaluate the performance of the classical periodogram approach to estimating the frequency of a sinusoid embedded in white Gaussian noise.

Since the problem of interest concerns the estimation of an unknown quantity in the presence of noise, it is useful to introduce several statistical concepts from estimation theory. Typically, it is desired to estimate the value of an unknown parameter θ from N noisy measurements of a quantity related to θ. An estimator is considered *unbiased* when the expected value of the estimate $\hat{\theta}$ equals the true value of the parameter:

$$E\{\hat{\theta}\} = \theta$$

If this condition does not hold, the estimator is termed *biased.* An estimator is considered *consistent* if it yields an estimate that asymptotically converges in probability to the true value. For a consistent estimator,

$$\lim_{N\to\infty} \Pr\{|\hat{\theta} - \theta| > \epsilon\} = 0$$

where Pr denotes probability and ϵ is an arbitrary small positive number. Both of these characteristics are generally thought to be desirable but, depending on the problem of interest, may or may not be required.

The benchmark by which the variance of a particular unbiased estimator can be evaluated is given by the Cramer–Rao lower bound (CRLB). As its name implies, the CRLB provides a lower bound on the variance of any linear or nonlinear unbiased estimator. Thus, given the variance of a particular unbiased estimator, the CRLB may be used to determine if other unbiased estimators might exist which exhibit smaller variance. Although other bounds on estimator variance exist, it is generally agreed that the CRLB is the easiest to compute, and hence finds extensive use (21). The CRLB for the scalar parameter θ is expressed in terms of the measurement vector $\boldsymbol{x}$ with

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{-E\left\{\dfrac{\partial^2 \ln p(\boldsymbol{x};\theta)}{\partial\theta^2}\right\}}$$

where $p(\boldsymbol{x};\ \theta)$ is the probability density function (PDF) of $\boldsymbol{x}$ given the parameter θ (22). The expectation is taken with respect to $p(\boldsymbol{x};\ \theta)$, which results in a function of θ. When the PDF is considered to be a function of the unknown parameter θ (with a fixed measurement vector x) it is called the likelihood function.

Although the CRLB may be computed for a specific estimation problem, there is no guarantee that an unbiased estimator exists which will equal the bound for all values of the unknown parameter. If such an estimator does exist, it is said to be *efficient.* An estimator is considered *asymptotically efficient* if its variance converges to the CRLB as the number of observations becomes large. *Maximum likelihood* (ML) estimators are known to be asymptotically efficient, and they can be constructed by computing the value of θ which maximizes the likelihood function. The performance of ML algorithms for large data records, along with the existence of an analytical approach to their derivation, makes ML algorithms very appealing. In practice, these algorithms can be computationally complex, such that other approaches may be preferred.

At high signal-to-noise ratio (SNR), the variance of ML estimators is typically very close to the CRLB. As the SNR is decreased, the CRLB and the estimator variance increase at the same rate. For nonlinear estimators, this behavior continues until a *threshold* is reached. Below this value of SNR, the variance of the estimator increases at a much faster rate than the CRLB. In a plot of the variance as a function of SNR, a knee will be seen at the threshold. Thus, the estimator threshold is frequently used as a metric to compare several estimators which have similar high-SNR characteristics. For maximum likelihood estimators, the threshold typically decreases as the size of the data window increases.

In addition to the factors described above, various algorithms for the estimation of a particular parameter may also be compared to one another with respect to computational considerations. Computational issues include: the algorithm's complexity as measured by the number of arithmetic operations (such as multiplications or arctan function calls) required for its implementation, the storage requirements of the algorithm, and the data window size required for satisfactory performance.

As a demonstration of the concepts described in this section, the classical approach to the estimation of the frequency of a single complex sinusoid with unknown amplitude and phase in the presence of white Gaussian noise is considered. The CRLB for the frequency estimate is given by

$$\mathrm{CRLB} = \frac{12}{T^2N(N^2-1)}\left(\frac{\sigma^2}{A^2}\right)$$

where A is the amplitude of the complex sinusoid, σ^2 is the power of the complex white Gaussian noise, T is the sampling period, N is the number of available data samples, and A^2/σ^2 is the SNR. To derive a ML estimator for this problem, the likelihood function L is computed in terms of the unknown frequency, amplitude, and phase. It has been shown that the likelihood function is given by

$$L(\omega, A, \theta) = 2A\mathrm{Re}[\exp(-j\theta)\exp(-j\omega t_0)X(\omega)]$$

where θ is the unknown phase, t_0 is the time corresponding to the first data sample, and $X(\omega)$ represents the discrete-time Fourier transform of the data sequence $x(nT)$ (23). The ML estimate of the sinusoid's frequency is the maximum of $L(\omega,\ A,\ \theta)$ with respect to ω over all values of A and θ, and it can be shown to correspond to the frequency which maximizes the periodogram $|X(\omega)|^2$. To implement this approach, it has been suggested that a coarse frequency estimate be obtained from the peak magnitude of the discrete Fourier transform (DFT) and that a more accurate result be obtained via an interpolation algorithm. The performance of the overall algorithm is significantly improved if the length of the original data sequence is increased by a factor of two or four by zero-padding, before the coarse DFT is computed. An accurate means of implementing the interpolation procedure with only three DFT points is presented in Ref. 24.

The ML algorithm for estimating the frequency of a single sinusoid in noise is conceptually simple, but computationally intensive. This is especially the case when the algorithm must be implemented under a real-time schedule in order to track a time-varying frequency. In the following section, additional approaches to the implementation problem are examined.

ALGORITHMS FOR INSTANTANEOUS FREQUENCY ESTIMATION

In this section, a selection of algorithms that have been suggested for IFE are summarized. This selection is not all-inclusive, and it is in fact concentrated in two areas: The first set of algorithms employ *weighted phase averaging* techniques, and the second set of algorithms are designed to function with extremely short data windows in high-SNR environments. Both sets of algorithms are designed for monocomponent signals. References to other approaches for estimating the instantaneous frequency of monocomponent signals, along with approaches for multicomponent signals, are provided.

Algorithms Employing Weighted Phase Averaging

One way to represent a constant amplitude, complex signal with time-varying frequency is to model the signal as a complex exponential with polynomial phase. From the Weierstass Theorem, it is known that all continuous phase functions can be approximated to any desired accuracy by a polynomial. The polynomial phase model is thus very general, in addition to being easily analyzed. When P coefficients are considered in the polynomial, the signal model is expressed by

$$z(n) = A \exp\left(j \sum_{p=0}^{P} c_{\mathrm{p}} n^{\mathrm{p}}\right) + \epsilon(n)$$

where A is the constant signal amplitude, n is the sampling index, c_{p} is the polynomial coefficient, and $\epsilon(n)$ is complex zero-mean white Gaussian noise. In the following discussion, algorithms which employ weighted phase averaging to estimate the coefficients of the polynomial phase model are examined. Before examining algorithms capable of estimating all P coefficients, less complex approaches corresponding to linear and quadratic phase models are summarized. The less sophisticated approaches are of interest due to their relatively undemanding computational requirements. In fact, it has been suggested that time-varying frequencies be tracked with sliding window implementations of these simpler algorithms.

A sinusoid with constant frequency can be represented by the polynomial phase model with a constant plus linear phase term. As discussed in the previous section, the ML estimate of the frequency of a single sinusoid embedded in white Gaussian noise is given by the peak of the periodogram. Unfortunately, the construction of the periodogram is computationally intensive, and other less complex approaches are desired. One such approach was suggested by Tretter (25). He considered the input data sequence to be modeled by

$$z(n) = A \exp(j(\theta + \omega n)) + \epsilon(n)$$

where θ is a constant phase and ω is the signal's angular frequency. The angular frequency is assumed to be bounded by $-\pi < \omega \leq \pi$. The noise power is given by σ^2, and the SNR is thus expressed as

$$\mathrm{SNR} = \frac{A^2}{\sigma^2}$$

Tretter showed that for SNR $\gg 1$, the data sequence can be approximated as

$$\begin{aligned} z(n) &\approx A \exp(j(\theta + \omega n + v(n))) \\ &\approx A \exp(j\phi(n)) \end{aligned}$$

where $v(n)$ is a real Gaussian white noise sequence with variance equal to 1/(2SNR). The impact of this approximation is that all of the information required to estimate the frequency ω is contained in the signal phase $\phi(n)$. Tretter suggested that the phase be estimated by unwrapping the sequence obtained from computing the arctan of $z(n)$. The frequency is then estimated via least squares or linear regression. For high SNR, this estimation scheme achieves the CRLB.

An alternate viewpoint to this problem was provided by Kay (26). He suggested that phase differences be employed rather than the phases themselves. The phase difference $\Delta(n)$ can be written as

$$\begin{aligned} \Delta(n) &= \phi(n+1) - \phi(n) \\ &= \omega + v(n+1) - v(n) \end{aligned}$$

The frequency estimation problem can then be expressed as the estimation of the mean of a colored Gaussian noise process. Kay showed that the ML frequency estimate for this problem is given by

$$\hat{\omega}_K = \sum_{n=0}^{N-2} w(n)\Delta(n)$$

where the total number of data samples available for processing is denoted by N, and $w(n)$ represents a parabolic weighting function given by

$$w(n) = \frac{1.5N}{N^2 - 1}\left[1 - \left(\frac{n - (0.5N - 1)}{0.5N}\right)^2\right]$$

Kay noted that if a uniform weighting is applied, the phase differences are merely averaged, and the variance of the estimate is increased by a factor equal to $N/6$ at high SNR. It was later shown that Kay's algorithm can be derived from Tretter's algorithm using summation-by-parts (27).

As is typical with nonlinear estimation methods, the variance of Kay's algorithm departs from the CRLB when the SNR is reduced below a threshold value. Kim noted that the threshold of Kay's algorithm occurs when the SNR drops below a value for which the phase noise approximation is valid. He suggested that the SNR of the signal be increased before the phase of the data samples is computed, by averaging K adjacent data samples. In this manner, the threshold is decreased, at a cost of a small loss in estimation performance and a decreased estimation range (28). For example, for data lengths greater than 24 and $K = 4$, Kim determined that his algorithm departs from the CRLB at high SNR by less than 0.2 dB. The threshold is reduced by a factor of 20 log(K) dB, and the estimation range is reduced by a factor of K.

In the frequency estimation work conducted by Rife and Boorstyn (23), it was noted that the angular frequency estimate of their algorithm (described in the previous section of this article) was biased whenever the angular frequency was close to zero or the sampling frequency. Similarly, the variance of Kay's estimator also significantly degrades when the angular frequency is close to these values. A means of overcoming this problem was proposed by Lovell and Williamson (29). They noted that the performance degradation is avoided if the weighting function is applied to the phase differences in a circular, rather than linear, fashion. For example, to compute the mean of a group of phases, they suggested that the phases first be expressed as unit magnitude phasors and that the argument of the sum of phasors then be computed. By incorporating these concepts into Kay's estimators, the sensitivity of the estimator variance with respect to angular frequency was significantly reduced.

The second coefficient relating to frequency in a polynomial phase model corresponds to frequency rate. Including this parameter μ, the signal model is written as

$$z(n) = A \exp(j(\theta + \omega n + \tfrac{1}{2}\mu n^2)) + \epsilon(n)$$

The frequency rate is assumed to be bounded by $-\pi < \mu \leq \pi$. This type of modulation is termed *linear frequency modulation* (LFM), and the corresponding signal is termed a *chirp* signal. Despite its simple form, this signal is utilized in many fields and is thus of significant interest.

A procedure to jointly estimate θ, ω, and μ for a chirp signal was suggested by Djuric and Kay (30). In this approach, the additive complex noise is modeled as real phase noise, as was the case in Refs. 25 and 26. However, a different technique is utilized to estimate the unambiguous phase sequence. First, two phase difference operations are implemented on the original data sequence, and the phase of the resulting data samples are computed with the arctan function. The sequence $d(n)$ is thus generated, where

$$d(n) = \mu + \Delta^2 w(n)$$

and $\Delta^2 w(n)$ denotes a colored noise sequence. An estimate $\hat{\phi}(n)$ of the unambiguous phase sequence $\phi(n)$ corresponding to the original data sequence is then obtained by twice integrating $d(n)$. The estimates of θ, ω, and μ are then jointly obtained from $\hat{\phi}(n)$. If only the frequency rate is desired, μ may be estimated directly from $d(n)$ in a similar fashion as ω was estimated in Ref. 26.

One shortcoming of the algorithm suggested by Djuric and Kay is its performance for large values of μ. When the magnitude of this parameter is close to its upper bound, errors occur in the phase unwrapping algorithm, and the performance of the estimator degrades. To overcome this effect, they suggested that a third phase difference operation be employed. However, this approach increases the probability of an outlier occurring due to differentiation of the phase noise, thereby degrading the unwrapping process and hence the estimation performance. An alternative solution to this problem was proposed by Slocumb and Kitchen (31). In their work, an iterative procedure is suggested in which the phase unwrapping and parameter estimation is conducted concurrently. A recursive least squares (RLS) algorithm (32) is employed to improve the phase unwrapping process, thereby removing the sensitivity of the threshold to the value of μ. For large values of μ, the threshold corresponding to Slocumb and Kitchen's approach is as much as 12 dB lower than the threshold of Djuric and Kay's algorithm.

The approaches presented above for chirp signals can be extended to estimate an arbitrary number of coefficients of the polynomial phase model. To prevent aliasing in a critically sampled signal, the polynomial coefficients must be bounded by

$$|c_{\mathrm{p}}| < \frac{\pi}{p!}$$

For the algorithm presented in Ref. 30, increasing the number of parameters to be estimated also increases the threshold of the algorithm.

Algorithms Employing Short Data Windows

In certain situations, it is reasonable to assume a very high SNR, even as high as 40 dB. It is then possible to obtain estimates of the instantaneous frequency of a monocomponent signal with only a few data samples. In this section, two computationally efficient algorithms are described which obtain accurate estimates of the instantaneous frequency with only four or five data samples. This feature is very desirable, because the instantaneous frequency estimate is thus highly localized in time.

Teager's energy operator was originally proposed as a means of quantifying the "energy" present in an harmonic oscillation (33). It has since been utilized to derive algorithms for instantaneous frequency estimation that are highly time localized. The discrete form of this operator is given by

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1)$$

where it is assumed that the sampling period is unity. Utilizing this operator, three different algorithms have been derived to estimate the instantaneous frequency and amplitude of a monocomponent AM-FM signal (34). The three algorithms are denoted DESA-1a, DESA-1, and DESA-2, and the associated instantaneous frequency estimation algorithms are expressed as

$$\omega_{1a}(n) = \arccos\left(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\right)$$
$$\omega_{1}(n) = \arccos\left(1 - \frac{\Psi[x(n) - x(n-1)] + \Psi[x(n+1) - x(n)]}{4\Psi[x(n)]}\right)$$
$$\omega_{2}(n) = \tfrac{1}{2}\arccos\left(1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]}\right)$$

The first algorithm requires four data points for its operation, and the remaining two algorithms require five data points. All three algorithms may be implemented with only a few multiplications per time step.

A second means of constructing highly time localized instantaneous frequency estimators is by symbolically expressing the roots of the predictor filter corresponding to a sinusoidal signal model in terms of the input data samples (35). Two forms of linear prediction have been examined for this application: the covariance method and the modified covariance method. For the modified covariance method, two estimators were derived in Ref. 35. The first estimator requires four data samples for its operation, and the second requires five data samples. The estimators are expressed in terms of the input data samples via

$$\omega_{MC4}(n) = \arccos\left(\frac{x(n-2)x(n-1) + 2x(n-1)x(n) + x(n)x(n+1)}{2(x^2(n-1) + x^2(n))}\right)$$

and

$$\omega_{MC5}(n) = \arccos\left(\frac{x(n-2)x(n-1) + 2x(n-1)x(n) + 2x(n)x(n+1) + x(n+1)x(n+2)}{2(x^2(n-1) + x^2(n) + x^2(n+1))}\right)$$

Utilizing the covariance method, a single estimator was derived that required five data samples for its operation:

$$\omega_{C5}(n) = \arccos\left(\frac{x(n-1)x(n) - x(n-2)x(n+1)}{x^2(n) - x(n-1)x(n+1) + x^2(n-1) - x(n-2)x(n)}\right)$$

These algorithms have been shown to require fewer computational operations per time step than the DESAs. The linear predictive algorithms also yield smaller mean and rms errors than the DESAs when simulated with signals having various amounts of amplitude and frequency modulation. The performance of linear predictive techniques with respect to the CRLB was investigated in Ref. 36.

Other Algorithms

The algorithms for IFE that were summarized in this section were concentrated in two general areas. Many other approaches for estimating the instantaneous frequency of monocomponent signals exist, such as the extended Kalman filter (37), the cross-power spectrum (38), and the discrete polynomial transform (39). Still other techniques are described in (19). For multicomponent signals, proposed approaches include adaptive notch filters (40), recursive least squares (41), cross-coupled digital phase-locked loops (42), and the periodic algebraic separation energy demodulation algorithm (43). Additional techniques are discussed in Ref. 43.

CONCLUSION

In this article, the connections between instantaneous frequency and time–frequency analysis were explored. Definitions for instantaneous frequency with respect to time–frequency distributions and to the analytic signal were provided. Measures of performance for estimators of instantaneous frequency were illustrated, such as the Cramer–Rao lower bound. Finally, a selection of algorithms that have been recently proposed for instantaneous frequency estimation were summarized.

ACKNOWLEDGMENTS

L. B. Fertig's research was supported by a GTRC Ph.D. fellowship.

J. H. McClellan's research was supported by the Joint Services Electronics Program under contract DAAH-04-96-1-0161.

BIBLIOGRAPHY

1. D. Gabor, Theory of communication, *Proc. IEE,* **93**: 429–457, 1946.
2. J. B. Allen and L. R. Rabiner, A unified approach to short-time Fourier analysis and synthesis, *Proc. IEEE,* **65**: 1558–1564, 1977.
3. L. R. Rabiner and R. W. Schafer, *Digital Signal Processing of Speech Signals,* Englewood Cliffs, NJ: Prentice-Hall, 1978.
4. A. Croisier, D. Esteban, and C. Galand, Perfect channel splitting by use of interpolation decimation tree decomposition techniques, *Int. Conf. Inf. Sci. Syst.,* 1976, pp. 443–446.
5. M. J. T. Smith and T. P. Barnwell, A procedure for designing exact reconstruction filter banks for tree structured subband coders, *Proc. ICASSP,* 1984, pp. 27.1.1–27.1.4.
6. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1989.
7. P. P. Vaidyanathan, *Multirate Systems and Filter Banks,* Englewood Cliffs, NJ: Prentice-Hall, 1993.
8. M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding,* Upper Saddle River, NJ: Prentice-Hall, 1995.
9. J. Morlet et al., Wave propagation and sampling theory, part 1: Complex signal and scattering in multilayered media, *Geophysics,* **47** (2): 203–221, 1982.
10. A. Grossman and J. Morlet, Decomposition of Hardy functions into square integrable wavelets of constant shape, *SIAM J. Math. Anal.,* **15** (4): 723–736, 1984.
11. I. Daubechies, Orthonormal bases of compactly supported wavelets, *Commun. Pure Appl. Math.,* **41**: 909–996, 1988.
12. S. Mallat, Multifrequency channel decompositions of images and wavelet models, *IEEE Trans. Acoust. Speech Signal Process.,* **37**: 2091–2110, 1989.
13. L. Cohen, *Time-Frequency Analysis,* Upper Saddle River, NJ: Prentice-Hall, 1995.
14. P. J. Kootsookos, B. C. Lovell, and B. Boashash, A unified approach to the STFT, TFD's, and instantaneous frequency, *IEEE Trans. Signal Process.,* **40**: 1971–1982, 1992.
15. L. Cohen, What is a multicomponent signal?, *Proc. ICASSP,* **5**: 113–116, 1992.
16. P. J. Loughlin and B. Tacer, Comments on the interpretation of instantaneous frequency, *IEEE Signal Process. Lett.,* **4** (5): 123–125, 1997.
17. B. Picinbono, On instantaneous amplitude and phase of signals, *IEEE Trans. Signal Process.,* **45**: 552–560, 1997.
18. A. Reilly, G. Frazer, and B. Boashash, Analytic signal generation—tips and traps, *IEEE Trans. Signal Process.,* **42**: 3241–3245, 1994.
19. B. Boashash, Estimating and interpreting the instantaneous frequency of a signal, Part I and II, *Proc. IEEE,* **80**: 519–568, 1992.
20. M. Sun and R. J. Sclabassi, Discrete-time instantaneous frequency and its computation, *IEEE Trans. Signal Process.,* **41**: 1867–1880, 1993.
21. S. M. Kay, *Fundamentals of Statistical Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1993.
22. H. L. Van Trees, *Detection, Estimation, and Modulation Theory,* Part I, New York: Wiley, 1968.
23. D. C. Rife and R. R. Boorstyn, Single-tone parameter estimation from discrete-time observations, *IEEE Trans. Inf. Theory,* **20**: 591–598, 1974.
24. B. G. Quinn, Estimating frequency by interpolation using Fourier coefficients, *IEEE Trans. Signal Process.,* **42**: 1264–1268, 1994.
25. S. A. Tretter, Estimating the frequency of a noisy sinusoid by linear regression, *IEEE Trans. Inf. Theory,* **31**: 832–835, 1985.
26. S. M. Kay, A fast and accurate single frequency estimator, *IEEE Trans. Acoust. Speech Signal Process.,* **37**: 1987–1990, 1989.
27. S. W. Lang and B. R. Musicus, Frequency estimation from phase differences, *Proc. ICASSP,* **4**: 2140–2144, 1989.
28. D. Kim, M. J. Narasimha, and D. C. Cox, An improved single frequency estimator, *IEEE Signal Process. Lett.,* **3** (7): 212–214, 1996.
29. B. C. Lovell and R. C. Williamson, The statistical performance of some instantaneous frequency estimators, *IEEE Trans. Signal Process.,* **40**: 1708–1723, 1992.
30. P. M. Djuric and S. M. Kay, Parameter estimation of chirp signals, *IEEE Trans. Acoust. Speech Signal Process.,* **38**: 2118–2126, 1990.
31. B. J. Slocumb and J. Kitchen, A polynomial phase parameter estimation phase unwrapping algorithm, *Proc. ICASSP,* **4**: 1867–1880, 1994.
32. S. Haykin, *Adaptive Filter Theory,* Englewood Cliffs, NJ: Prentice-Hall, 1991.

33. J. F. Kaiser, On a simple algorithm to calculate the "energy" of a signal, *Proc. ICASSP,* **1**: 381–384, 1990.
34. P. Maragos, J. F. Kaiser, and T. F. Quatieri, On separating amplitude from frequency modulations using energy operators, *Proc. ICASSP,* **2**: 1–4, 1992.
35. L. B. Fertig and J. H. McClellan, Instantaneous frequency estimation using linear prediction with comparisons to the DESA's, *IEEE Signal Process. Lett.,* **3** (2): 54–56, 1996.
36. S. W. Lang and J. H. McClellan, Frequency estimation with maximum entropy spectral estimators, *IEEE Trans. Acoust. Speech Signal Process.,* **28**: 716–724, 1980.
37. K. Nishiyama, A nonlinear filter for estimating a sinusoidal signal and its parameters in white noise: On the case of a single sinusoid, *IEEE Trans. Signal Process.,* **45**: 970–981, 1997.
38. S. Umesh and D. Nelson, Computationally efficient estimation of sinusoidal frequency at low SNR, *ICASSP,* **5**: 2797–2800, 1996.
39. S. Peleg and B. Friedlander, Signal estimation using the discrete polynomial transform, *Proc. ICASSP,* **4**: 424–427, 1993.
40. G. Li, A stable and efficient adaptive notch filter for direct frequency estimation, *IEEE Trans. Signal Process.,* **45**: 2001–2009, 1997.
41. P. Tichavsky and P. Handel, Efficient tracking of multiple sinusoids with slowly varying parameters, *Proc. ICASSP,* **3**: 1993, pp. 368–371.
42. J. N. Bradley and R. L. Kirlin, Phase-locked loop cancellation of interfering tones, *IEEE Trans. Signal Process.,* **41**: 391–395, 1993.
43. B. Santhanam, *Multicomponent AM–FM energy demodulation with applications to Signal Processing and Communications,* Ph.D. dissertation, Georgia Institute of Technology, Atlanta, GA, 1997.

L. B. FERTIG
J. H. MCCLELLAN
Georgia Institute of Technology

FREQUENCY RESPONSE. See TRANSFER FUNCTIONS; DIODES.