Signals and Communication Technology

Rainer E. Gruhn
Wolfgang Minker
Satoshi Nakamura

Statistical Pronunciation Modeling for Non-Native Speech Processing

Springer

Signals and Communication Technology

For further volumes:
http://www.springer.com/series/4748

Rainer E. Gruhn · Wolfgang Minker
Satoshi Nakamura

Statistical Pronunciation Modeling for Non-Native Speech Processing

Springer

Rainer E. Gruhn
Harman/Becker Automotive
Systems GmbH
Söflinger Str. 100
89077 Ulm
Germany
e-mail: rrgruhn@hotmail.com

Wolfgang Minker
Universität Ulm
Fak. Ingenieurwissenschaften und
Elektrotechnik
Abt. Informationstechnik
Albert-Einstein-Allee 43
89081 Ulm
Germany
e-mail: wolfgang.minker@uni-ulm.de

Satoshi Nakamura
Department of Acoustics and
Speech Research
NICT-ATR
Hikaridai 2-2-2
Keihanna Science City
Kyoto 619-0288
Japan
e-mail: satoshi.nakamura@nict.go.jp

ISSN 1860-4862

ISBN 978-3-642-19585-3 e-ISBN 978-3-642-19586-0

DOI 10.1007/978-3-642-19586-0

Springer Heidelberg Dordrecht London New York

Cover design: eStudio Calamar, Berlin/Figueres

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Automatic speech recognition systems are increasingly applied for modern communication. One example are call centers, where speech recognition based systems provide information or help sorting customer queries in order to forward them to the according experts. The big advantage of those systems is that the computers can be online 24 h a day to process calls and that they are cheap once installed.

Travelers can find speech-to-speech translation systems commercially available, either via cellular phone or based on hand-held devices such as PDAs. These translation systems typically provide translation from English to Japanese (or in case of American military systems, to Arabic) and back. Pressing a push-to-activate button starts the recording of an utterance, which is recognized and translated. The translated sentence is then played to the communication partner with a text-to-speech (TTS) module.

Speech control is also a common feature of car navigation devices, for command and control purposes as well as destination input. Undeniably, speech control increases comfort. It is also undisputed that speaking a telephone number whilst driving rather than typing it on the tiny cellular phone keyboard is an important safety measure, as the driver can still look and concentrate on traffic without too much distraction by the input procedure. Speech controlled music players will enter the mass market soon.

Moreover, in many of these applications of speech recognition, there are cases when the speaker is not expressing himself in his mother tongue. Customers of call centers are people who are not necessarily citizens of the country where the service is offered nor have they mastered the local language to a native level of proficiency. Translation systems for travelers as sold e.g. in Japan assume that most foreign dialog partners the owner of a Japanese–English translation system encounters speak English, native or as a second language. Car navigation systems must offer drivers the possibility to set a destination abroad, therefore good systems have to support speech input of foreign place names. Finally, given the strong internationalization of the music market, speech controlled music players must cope with non-native speech input, e.g. for English language songs in Germany or Spanish language songs in the United States.

A speech recognition system recognizes words as a sequence of phonemes defined in a pronunciation dictionary. These sequences do not entirely match non-native speaker utterances as they deviate from the standard pronunciations of words, inserting and omitting sounds as typical for the phonetic contexts of the native language. They especially generate different sounds that are more familiar from the speakers mother tongue but do not fully match the phonetic inventory of the language the speaker has not fully mastered. For both humans and machines, these deviations are a big hurdle to understand what a non-native speaker says. By listening to accented speech for some time, humans can learn the specific accent patterns and adapt to them to some extend.

The target of this research is to provide a method that adjusts an automatic speech recognition system so that it can recover some of the errors caused by non-native pronunciation. We relax the pronunciation dictionary constraints for recognition of non-native speech. Then by training on a non-native speech sample, we factor in the specific pronunciation error patterns of each accent without attempting to represent them explicitly.

The authors would like to thank Seiichi Yamamoto for providing the opportunity to do research at ATR. Furthermore we thank Konstantin Markov for valuable discussions and helpful suggestions and Nobuaki Minematsu and Frank Soong for analyzing this work and asking questions.

The first author is especially grateful to Heinrich Niemann, Elmar Nöth, Yoshinori Sagisaka and Harald Singer for introducing him to the marvels of automatic speech recognition and its practical application.

We acknowledge the valuable insights the research of Norbert Binder, Tobias Cincarek, Bi Lige, Martin Raab and Wenzel Svoyanovsky have contributed to this work.

Thanks go to Nick Campbell and JST/CREST ESP for granting access to the expressive speech database.

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology (NICT) of Japan entitled, “A study of speech dialog translation technology based on a large corpus”.

Contents

Chapter 1
Introduction

1.1 This Book in a Nutshell

Recognizing the speech of non-native speakers is a challenge for automatic speech recognition systems. Non-native speech recognition is an important issue in applications such as speech-to-speech translation devices or speech dialog systems.

The pronunciation of non-natives deviates from native speech, leading to insertions, deletions and especially substitutions of sounds. The accent pattern depends mainly on the mother tongue of the speaker, phoneme confusion matrices can provide an illustration of those patterns.

There are two approaches to obtain knowledge about non-native pronunciation: By expert analysis of a language, or by data-driven methods such as phoneme recognition and comparison of the result to a native transcription. Only the data-driven method is capable of handling multiple accents automatically.

We designed, collected and transcribed the ATR non-native English database consisting of 96 speakers of five accent groups, Mandarin Chinese, French, German, Italian and Japanese English. The data includes hotel reservation dialogs, number sequences and phonetically compact sentences. This database is one of the largest non-native databases, and has been made available at ATR to promote non-native speech research and computer-assisted language learning system development. It has been recorded with special care about countering anxiousness, a special problem for non-native speech collections. It does not suffer from noise or exotic vocabulary like military databases and does not have the spontaneous speech effects of presentation or meeting transcriptions.

Native speakers with English teaching experience have rated the skills of the speakers and marked mispronounced words. A rater calibration helped increasing inter-rater consistency. About 9.3% of the words have been marked as mispronounced, some were difficult for all accents, others accent-specific errors.

R.E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*,
Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_1,

These properties make the database most interesting for researchers in the field of non-native speech and we hope it will help to make non-native speech publications more comparable.

A common way to represent the non-native pronunciation variations is in the form of confusion rules. We experimented with phoneme confusion rules extracted from the training data set of the collected database. We examined their usefulness on for phoneme lattice processing. When applying them on a pronunciation dictionary, we found a tradeoff between considering pronunciation variations and causing additional confusions: Adding some pronunciation variants helps, but adding too many impairs recognition performance.

To automatically score the skill of a non-native speaker and detect mispronounced words, we evaluated several feature types. To calculate a sentence-level pronunciation score, the phoneme likelihood ratio and the phoneme accuracy are the most reliable features. A class recognition rate of 86.6% was achieved. To identify mispronounced words, the phoneme confusion ratio and the word likelihood contributed well. Weighting false alarm errors higher, experiments showed a 72% recognition rate, a reliability comparable to the human experts.

The core proposal of this book is a fully statistical approach to model non-native speakers' pronunciation. Second–language speakers pronounce words in multiple different ways compared to the native speakers. Those deviations, may it be phoneme substitutions, deletions or insertions, can be modeled automatically with the proposed method.

Instead of fixed phoneme sequences as pronunciation baseforms, we use a discrete HMM as a word pronunciation model. It is initialized on a standard pronunciation dictionary, taking into account all baseform pronunciations. One HMM is generated per word in the dictionary, with one state per phoneme in the baseform pronunciation.

Non-native training data is segmented into word-length chunks, on which phoneme recognition is performed. The probability distributions of the HMMs are trained on these phoneme sequences. Evaluations showed that for non-native speech, monophone acoustic models perform better than context-dependent models, a tendency different from native speech. The likely reason is that the coarticulation is different for non-native speech and that context-dependent models are trained too closely to the specific acoustic context to deal with non-nativeness effects.

To apply the models, both an N-best word level recognition and a utterance–level phoneme recognition of the test data are required. A pronunciation score is calculated by performing a Viterbi alignment with the HMM dictionary as model and the phoneme sequence as input data. This score is a measure how well the phonemes match with the pronunciation of the word sequence modeled by the HMM. The n-best list is resorted according to pronunciation score, the hypothesis with the highest score is selected as recognition result.

On the collected database, we conducted experiments to verify the applicability of the proposed method. For each non-native speaker group, pronunciation HMMs are trained and evaluated on a test set of non-native English in the regarding accent. The approach improves the recognition accuracy for all speaker groups, the best results were achieved for Mandarin Chinese speakers with a relative 11.93% gain.

Experiments to combine the ”pronunciation score” calculated with the pronunciation HMMs together with language model score or acoustic model score were conducted as well. Factoring in a language model showed to be an important contribution, whereas the acoustic score does not contain additional information. Experiments further showed that already rescoring with the initial models increased performance to some extend, showing that the general relaxation of pronunciation constraints from the baseform dictionary is already helpful.

1.2 Contribution of this Research

The contributions of this work on statistical language modeling of non-native pronunciations to the research community are:

- Concept design, implementation and evaluation of a novel approach to handle the pronunciation variations of non-native speech in an implicit and statistical way using discrete hidden Markov models (HMM) as statistical dictionary. The proposed models are proven to be effective in increasing recognition rates of non-native speakers, independent of accent and without expert knowledge necessary.
- Creation of a large database of non-native speech of around 100 speakers from 5 different accent groups, recorded under the most appropriate conditions and including human expert pronunciation ratings.

While general pronunciation networks have already been proposed, modeling non-native pronunciations with word-level HMMs and applying them for rescoring is an original approach. Our evaluations show that they are effective and can allow for any pronunciation variation, regardless of whether these variations have been previously observed in the training data or not. Our method does not decrease recognition performance due to additional confusions, which seems to be a significant problem for the most common method of phoneme confusion rule generation. As it is a fully data-driven approach it may be applied to any accent, i.e. pair of native and non-native language, without any need for expert knowledge about the accent.

The outline of this book is as follows: First, we present the basics of automatic speech recognition, followed by a brief introduction to non-native speech and its challenges in Chap. 3. A literature survey describes the state of the art for non-native speech recognition in Chap. 4. A significant contribution of this work is the collection of a large non-native speech database, whose properties are shown in Chap. 5. After describing experiments on rule-based non-native speech processing, multilingual codebooks and automatic pronunciation skill scoring in Chap. 6, we propose and evaluate HMMs as a statistical lexicon for non-native speech recognition in Chap. 7. The document ends with a conclusion, critically discussing the work and giving future directions. The appendix contains additional information about the ATR non-native database, including phoneme confusion matrices, a hotel reservation dialog sample, speaker properties and the human expert rating procedure.

Chapter 2
Automatic Speech Recognition

Automatic speech recognition (ASR) systems convert speech from a recorded audio signal to text. Humans convert words to speech with their speech production mechanism. An ASR system aims to infer those original words given the observable signal. The most common and as of today best method is the probabilistic approach. A speech signal corresponds to any word (or sequence of words) in the vocabulary with a certain probability. Therefore, assuming a word x or word sequence X was spoken, we compute a score for matching these words with the speech signal. This score is calculated from the acoustic properties of speech sub-units (phonemes in the acoustic model), linguistic knowledge about which words can follow which other words. Including additional knowledge as the pronunciation score proposed in this work has also shown to be helpful. Finally, we sort the possible word sequence hypotheses by score, and pick the hypothesis with the highest score as recognition result.

The outline of a typical speech recognition system is shown in Fig. 2.1. The process of speech recognition can be divided into the following consecutive steps.

- pre-processing (which includes speech/non-speech segmentation,)
- feature extraction,
- decoding, the actual recognition employing an acoustic and language model as well as a dictionary,
- result post-processing.

In the following, these steps will be described in more detail.

2.1 Relevant Keywords from Probability Theory and Statistics

In order to understand automatic speech recognition, it is helpful to briefly review some key concepts from general probability theory. For further reading, we refer to [Huang 01, Bronstein 07].

R.E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*,
Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_2,

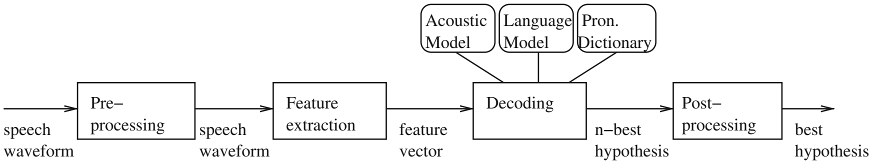
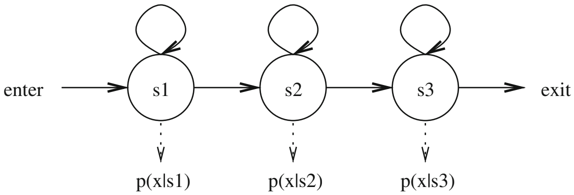


Fig. 2.1 Schematic outline of a typical speech recognition system

Table 2.1 Example for a discrete probability distribution

	$X = a$	$X = b$	...	$X = x$
P(X)	0.1	0.07	...	0.09

Fig. 2.2 A three-state hidden Markov model



2.1.1 Discrete and Continuous Probability Distribution

A discrete random variable X describes random experiments. An example is rolling a dice, where $X \in [1...6]$. The probability distribution $P(X = i)$ can be given as shown in Table 2.1. That means, p(x) is a non-parametric discrete distribution.

In the example shown in Table 2.1, X was a natural number. For the continuous case with $X \in \mathbb{R}$, we have a probability density function $p(X = x)$ with $\int_{-\infty}^{\infty} p(X)dX = 1$. The true data density function is estimated with a parametric curve. A mixture of N Gaussians is a good approximation to most densities:

$$p(X) = \sum_{i=1}^{N} \frac{1}{(\sqrt{2}\pi) \mid \sigma_i \mid} \exp\left[-\frac{(X-\mu_i)^2}{2\sigma_i^2}\right] \tag{2.1}$$

with μ_i the mean and σ_i the variance of the ith Gaussian.

2.1.2 A Hidden Markov Models

A hidden Markov model (HMM) is a statistical model for a Markov process with hidden parameters. Figure 2.2 shows an example for a three-state HMM. Each state s_i has a probability density $p_i, p(x|s_i)$ more precisely states the probability density for the acoustic observation x for the state s_i. The three states s_1, s_2 and s_3 together form the HMM for the word S. The observations X could be acoustic observations or, in case of the pronunciation models, discrete phoneme symbols.

An HMM can be characterized by two sets of parameters: The transition matrix A of the probabilities $a_{s_{t-1}s_t}$ to go from one state to another (including self-loops) and the output probabilities B characterized by the densities.

2.1.3 Estimating HMM Parameters

HMMs are trained on data samples with the forward-backward or Baum-Welch algorithm [Baum 66], which is similar to the EM-algorithm [Schukat-Talamazzini 95]. There is no analytical method to determine the HMM parameters, the Baum-Welch algorithm performs an iterative estimation, which monotonously improves the HMM parameter set $\Phi = (A, B)$. [Huang 01] describes the algorithm in four steps as follows:

1. Initialization: Choose an initial estimate Φ
2. E-Step: Compute the auxiliary function $Q(\Phi, \Phi')$ based on Φ
3. M-step: Compute Φ' to maximize the auxiliary Q-function.
4. Iteration: Set $\Phi = \Phi'$, repeat from step 2

with Q defined as

$$Q(\Phi, \Phi') = \sum_{S} \frac{P(X, S|\Phi}{P(X|\Phi)} \log P(X, S|\Phi') \tag{2.2}$$

where

$$P(X, S|\Phi) = \prod_{t=1}^{T} a_{s_{t-1}s_t} p_{s_t}(x_t) \tag{2.3}$$

The initial parameters play an important role for this algorithm. There is no guarantee that the algorithm will converge, the ideal number of iterations is typically determined heuristically.

HMM models are trained on data, for example an acoustic model on a speech database. If the corpus includes the data of a sufficient number of speakers, it can be assumed to be general, i.e. the model will acceptably represent the properties of speakers not observed in the training database. Such a model is called speaker-independent.

If the model is trained on a group of specific speakers, it is considered group-dependent; for example a model trained on speakers from an accent group is accent-dependent. It is theoretically possible to train a speaker-dependent model on the data of a single speaker. However, such models are less reliable as there is a high chance that many relevant data items (such as phonemes in certain contexts) have not been uttered by that speaker. There are algorithms to adapt a general acoustic model to the properties of a specific speaker, such as the maximum a posteriori (MAP) adaptation. Such models will have increased recognition performance for the regarding subject but less match other speakers.

The MAP estimate to optimize the parameters Φ' of an HMM can be expressed as follows [Huang 01]:

$$\Phi' = \mathrm{argmax}_{\Phi}[p(\Phi|X)] = \mathrm{argmax}_{\Phi}[p(X|\Phi), p(\Phi)] \tag{2.4}$$

This equation can be solved with the EM algorithm, with the Q-function defined as:

$$Q_{MAP}(\Phi, \Phi') = \log p(\Phi') + Q(\Phi, \Phi') \tag{2.5}$$

2.2 Phonemes

Speech is composed of certain distinct sounds. A phoneme is defined as the smallest unit of speech that distinguishes a meaning. Phonemes are characterized by the way they are produced, especially:

- place of articulation,
- manner of articulation,
- voicing.

For each language, there is a specific set of phonemes. There are several notations how to transcribe these phonemes. The most notable is the International Phonetic Alphabet (IPA) [Association 99] which was developed by the International Phonetic Association beginning in 1888 with the goal of describing the sounds of all human languages. IPA consists of many symbols that are not normal Latin characters, i.e. not included in the standard ASCII codepage. This makes it inconvenient to use them on computers.

Several systems have been proposed to transcribe phonemes with ASCII symbols, with SAMPA [Wells 95] and ARPAbet [Shoup 80] being the most common. For these phonetic alphabets mapping tables exist that convert from one phonetic alphabet to the other, there is no fundamental difference between the alphabets that could influence experimental results. In this work phonemes are given in the ARPAbet notation, as the acoustic model is trained on the WSJ database and the pronunciation dictionary for that data is provided in ARPAbet notation.

For acoustic models in speech recognition, the units can be phonemes or units considering phonemes and their acoustic contexts. Units considering only the left or right context are called biphones, if they consider left and right context, they are called triphones.

2.3 Prosody

Besides the phonemes that carry the textual content of an utterance, prosodic information [Noeth 90] gives valuable support to understand a spoken utterance. In short, prosody is the rhythm, stress and intonation of continuos speech, and is expressed in pitch, loudness and formants. Prosody is an important mean of conveying non-verbal information.

Fujisaki [Fujisaki 96] considers two separate aspects of prosody, the "concrete aspect"—defining prosody in physical term, and the "abstract aspect' '—defining prosody as influence to linguistic structure.

concrete aspect: phenomena that involve the acoustic parameters of pitch, duration, and intensity

abstract aspect: phenomena that involve phonological organization at levels above the segment

Prosody in speech has both, measurable manifestations and underlying principles. Therefore the following definition is appropriate:

Prosody is a systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production. Its realization involves both segmental features of speech, and serves to convey not only linguistic information, but also paralinguistic and non-linguistic information.

The individual characteristics of speech are generated in the process of speech sound production. These segmental and suprasegmental features arise from the influence of linguistic, paralinguistic, and nonlinguistic information. This explains the difficulty of finding clear and unique correspondence between physically observable characteristics of speech and the underlying prosodic organization of an utterance.

Linguistic information: symbolic information that is represented by a set of discrete symbols and rules for their combination i.e. it can be represented explicitly by written language, or can be easily and uniquely inferred from the context.

Paralinguistic information: information added to modify the linguistic information. A written sentence can be uttered in various ways to express different intentions, attitudes, and speaking styles which are under conscious control of the speaker.

Nonlinguistic information: physical and emotional factors, like gender, age, happiness, crying, … which cannot be directly controlled by the speaker. These factors are not directly related to (para-) linguistic contents, but influence the speech anyway.

Prosodic characteristics are typically expressed in several types of features, which can serve as basis for automatic recognition. The most prominent of those features are duration, loudness, pitch and glottal characteristics.

2.3.1 Duration

Utterances can be lengthened or shortened; the relative length carries prosodic information. For example, [Umeno 03] shows that short non-verbal fillwords show affirmation, whereas lengthened fillwords express disagreement.

2.3.2 Power

The signal power or loudness of an utterance is another important prosodic feature. In German and English the intensity often marks or emphasizes the central information of a sentence. Without this information, spontaneous speech could be ambiguous and easily misunderstood. The loudness is measured by the intensity of the signal energy.

2.3.3 Pitch

At the bottom of the human vocal tract are the vocal cords, or glottis. For unvoiced speech, the glottis remains open, for voiced speech it opens and closes periodically. The frequency of the opening is called the fundamental frequency or pitch. It can be calculated from the spectrum [Kiessling 92] and its contour over the utterance reveals several information. E.g. in Mandarin Chinese, the F0 carries phonetic/lexical information, and in English or German, the pitch specifies a question by a final fall-rise pattern [Sun 06, Waibel 88].

2.3.4 Glottal Characteristics

Physiological voice characteristics also contribute to convey non-verbal information. The glottis is the vocal cord area of the human articulatory system and is most commonly known for creating voicing in pronunciation by opening and closing periodically. The interpretation and extraction of glottal characteristics directly from the waveform without the need of special recording equipment is described in literature [Hanson 97].

2.4 Speech to Text

In this section, we describe the steps from speech to text, beginning with recording and pre-processing, over feature calculation to decoding and finally rescoring.

2.4.1 Pre-processing

Speech is recorded with a microphone and the signal is discretized with a sampling frequency of e.g. 16 kHz. The Shannon sampling theorem states that a bandwidth limited signal can be perfectly reconstructed if the sampling frequency is more than double of the maximum frequency. That means that in the sampled data, frequencies up to almost 8 kHz are constituted correctly. While this is not the total frequency range of human speech, it is more than double of what is transmitted over telephone networks. These are typically limited to the 5 Hz–3.7 kHz range, and has shown in research to be sufficient for speech recognition applications. It is possible to remove frequencies below 100 Hz with a high-pass filter as they tend to contain noise but can be considered of little relevance for speech recognition.

An important part of pre-processing is also speech/non-speech segmentation. As speech recognition systems will classify any sound to any phoneme with some (even if very low) probability, background noise can cause insertions of phonemes or words into the recognition result if the noise resembles the parameters of a phoneme model better than those of a silence model. Such insertions can be reduced by removing areas from the speech signal between the start of the recording and the point of time when the user starts to speak, and after the end of the utterance. This segmentation process is also called end point detection.

Signal energy based algorithms have been available for a long time [Rabiner 75, Reaves 93, Junqua 94]. When the signal energy exceeds a given threshold, the start of a speech segment is detected. When the signal drops below a certain threshold, the speech segment ends. As there are phonemes with low signal energy and short pauses between words, this algorithm must be enhanced by time windowing or additional prosodic features [Noeth 90, Gruhn 98] such as voicedness to be reliable. Another common approach is based on Gaussian mixture models [Binder 01]. Video signals can also be very helpful in detecting speaking activity [Murai 00, Nishiura 01a, Nishiura 01b, Nishiura 02a, Nishiura 02b].

2.4.2 Feature Extraction

To calculate features, acoustic observations are extracted over time frames of uniform length. Within these frames, the speech signal is assumed to be stationary. The length of these frames is typically around 25 ms, for the acoustic samples in this window one multi-dimensional feature vector is calculated. The time frames are overlapping and shifted by typically 10 ms. On the time window, a fast Fourier transformation is performed, moving into the spectral domain.

Human ears do not perceive all frequency bands equally. This effect can be simulated with band-pass filters of non-uniform frequency band widths. Until 500 Hz, the width of the filters is 100 Hz, after that it increases logarithmically. The filter center frequencies are defined in the so called Mel scale. The spectrum is decorrelated with a discrete cosine transformation. Of the resulting coefficients, the first coefficients carry the most significance. Therefore only the first e.g. ten coefficients are selected as feature vector. The resulting features are called Mel cepstra, commonly abbreviated as MFCC. Usually the normalized energy is appended to the feature vector.

A common further feature processing step is cepstral mean subtraction (CMS). The goal of CMS is to remove the effects of a linear filter. For example the microphones for recording the training data and the microphone during testing are often different, CMS can contribute to recover from this effect. The average of the features is calculated for every utterance, and subtracted from each feature vector. As a result, the features during both test and training have a mean of zero.

Information lies not only in the feature vector itself, but also in the temporal change. There are two common approaches how to capture this information:

- create a supervector concatenating consecutive feature vectors,
- append the derivatives and second derivatives to the feature vector.

The first method will lead to a vector of very high dimensionality that must be projected down to a lower dimension with algorithms like principal component analysis or linear discriminant analysis [Vasquez 08]. Dimensionality reduction can also be applied to a feature vector with derivatives.

2.4.3 Decoding

Decoding is the process to calculate which sequence of words is most likely to match to the acoustic signal represented by the feature vectors. For decoding three information sources must be available:

- an acoustic model with an HMM for each unit (phoneme or word,)
- a dictionary, typically a list of words and the phoneme sequences they consist of,
- a language model with word or word sequence likelihoods.

A prerequisite for decoding is to know which words can be spoken. Those words are listed in the dictionary, together with the according phoneme sequence. The acoustic model typically has a probability density function that is a mixture of Gaussians and gives a likelihood for each observed vector $p(x|w)$.

A language model is not an absolute requirement for decoding but increase word accuracy [Tanigaki 00]; in some cases like a credit card recognition system with a vocabulary consisting of the numbers 0–9, it can be acceptable to consider all words equally likely. Language models are typically fixed grammars or n-gram models. A 1-gram model lists words and their likelihoods, a 2-gram model lists words and their likelihood given a preceding word and so on, providing the word probability $p(w)$.

During decoding, we search for the word(s) w^* that fits best to the observation X, as given in this fundamental equation:

$$w^* = \mathrm{argmax}_w(p(X|w)p(w)) \tag{2.6}$$

with $p(w)$ Coming from the language model and $p(X|w)$ calculated from the sequence of phonemes in the word as defined by the dictionary:

$$p(X|w) = \text{argmax}_s(\prod_j(p(x|s_j)p(s_j))) \tag{2.7}$$

In theory, it is also necessary to consider $p(x)$, but as this term is the same for all competing hypotheses, it can be neglected.

As the space of possible state sequences is astronomically large, it is not possible to calculate the probabilities of all existing paths through the state network: for T observations and N states, the complexity is $O(N^T)$. To find the most likely sequence of hidden states, the Viterbi search algorithm [Viterbi 79] is employed. It can be summarized into four steps [Jelinek 97, Huang 01] as follows: To find the optimal state sequences, find the maximizing sequence $s = s_1, \ldots, s_j, \ldots, s_{i-1}, s_i$ whose probability is $V_t(i)$ to generate the observation X_t at time t and ends in state i.

- Initialization: Set $V_0(s_0) = 1$
  $V_1(s) = \max_{s'} p(x_1, s|s')V_0(s') = p(x_1, s|s_0)$
- Induction: $V_t(s) = \max_{s'} Vt - 1(s')p(x_{t-1}, s|s')$ i.e. for each state s_j at time t keep only one path that leads to this state and discard all paths with lower probability.
- Termination: Find the state s^* at the end of the state sequence s_i where $V_t(s)$ is maximal.
- Traceback: Trace back from this state s^* to the initial state along the remaining transitions. The states along this path constitute the most likely state sequence $S^* = (s_1^*, s_2^*, \ldots, s_T^*)$.

The complexity of the Viterbi algorithm is only $O(N^2T)$. The list of all permitted paths in the state network is called the lattice.

The dictionary contains a list of all words defined in a recognition scenario and the phoneme sequence (and thereby HMM phoneme model sequence) of each word. If the number of words is very small and those words are acoustically different, very high speech recognition accuracies can be achieved. The larger the dictionary, the more confusions are possible, leading to decreasing recognition rates. The confusability depends not directly on the number of words, but on the number of entries, i.e. pronunciation alternatives. Words can be spoken differently even in native language, such as the digit 0 in English as /zero/ or /o/. Hitherto pronunciation dictionaries typically have more entries than words. Other than specialized recognition tasks such as digit recognition, most large vocabulary speech recognition systems have dictionary sizes of several 1,000 words, dictation systems can reach several 10,000 words. In a real-world system it is not always possible to predict which words a user might say. The larger the dictionary, the less words fail recognition as out of vocabulary, but adding many new words with similar phonemes leads to additional confusions and decreases recognition rates. The language model is also affected by vocabulary size: The smaller the number of words, the sharper a language model can describe the permitted word sequences, leading to a direct relation between number of words and recognizer performance.

2.4.4 Post-processing

The result of the Viterbi search is not a single sequence of words, but a list of all possible hypotheses sorted by total score. In practice, this number is usually limited to the five or ten best hypotheses, the so-called n-best list. Rescoring this list by employing additional sources of information is a common method to improve the recognition accuracy of the top-scoring result.

A possible source of information is a higher-order language model. As they require much more resources than unigram or bigram models, both in terms of memory and computation cost, combinations of bigram model for decoding followed by trigram model based rescoring are common.

In this work, we provide additional information about the pronunciation with pronunciation models and apply them with a rescoring algorithm.

2.5 Applying Speech Recognition

In order to apply a speech recognition system, it is necessary to judge the performance and to put it in a system environment with other related modules.

2.5.1 Evaluation Measures

The most common measures to evaluate the performance of a speech recognition system are correctness, accuracy and error. These measures can be calculated on phoneme and word level. There are three types of mistake a speech recognition system can make:

- Substitution: At the position of a unit (word or phoneme), a different unit has been recognized.
- Deletion: In the recognition result a unit is omitted.
- Insertion: The result contains additional units than were not spoken.

The type of error is determined by comparison with the correct transcription. Some transcriptions contain information about noises in the speech signal. Even though human noises can contain relevant information [Svojanovsky 04], omitting those non-verbal sounds is not counted as recognition mistake. Likewise, an acoustic model may contain special models to handle human noises such as breathing and lip smacks or environment noises; additional noises in the recognition result are not counted as insertion.

Correctness is defined as:

$$\mathrm{Corr} = \frac{N - D - S}{N} \tag{2.8}$$

with N being the number of units in the correct transcription, D the number of deletions and S the number of substitutions. The accuracy is similar to the correctness, but additionally takes the number of insertions I into account:

$$Acc = \frac{N - D - S - I}{N} \tag{2.9}$$

Both measures are commonly given in percentage notation. Finally the error rate is the sum of all types of errors divided by number of words in the transcription. The error rate is closely related to the accuracy, usually only one of the two numbers is given.

$$Err = \frac{D + S + I}{N} = 100\% - Acc \tag{2.10}$$

In case the units are words, the error rate is commonly abbreviated WER, standing for word error rate.

To calculate correctness and accuracy, the transcription is compared to the highest scoring recognition result. In some cases it is necessary to determine the best matching path among all possible paths in the recognition network to measure the best theoretically achievable result. Such an extended evaluation yields the so-called network correctness or network accuracy.

2.5.2 Speech Dialog Systems

In real-world applications, speech recognition are not employed stand-alone, but embedded in an larger architecture that also involves other modules. The most perceivable module other than speech recognition is speech synthesis or text-to-speech (TTS). A TTS system receives a text string (and optionally phonetics) and generates an audio file, which is played to the user as feedback.

Recognition and TTS are the user communication interfaces of a speech dialog system (SDS). Depending on the application, an SDS also includes one or several backend modules. Most important is a dialog manager, which implements the dialog strategy [Minker 02b], i.e.:

- keep track of the information the user has already uttered,
- know which data must be available in order to fulfil a given task,
- trigger database queries,
- decide the next dialog step, especially feedback or question to the user.

Frequently employed other modules include a database access module or a natural language understanding module [Minker 98b] which extracts keywords and data from spontaneous utterances. Typical applications for SDSs are information retrieval services, such as travel assistance systems (e.g. [Seneff 98]) for hotel reservation, ticket booking etc., with a telephony or information kiosk interface.

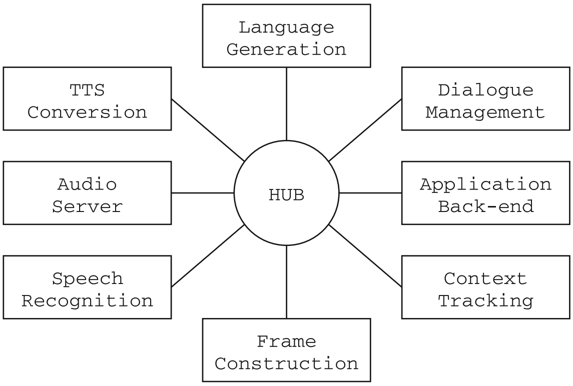


Fig. 2.3 The DARPA Communicator [Seneff 98]

Also the control of a car navigation system can base on an SDS, keeping track of previously input information and guiding the user during navigation destination input.

There are several well-known systems to combine the modules to a SDS. The "DARPA Communicator", also known by the name of its reference implementation "Galaxy", is a modular architecture developed at MIT [Seneff 98]. A schematical layout is shown in Fig. 2.3. It was designed for speech recognition based information retrieval systems. It consists of a central hub which interfaces between servers like audio server, dialog manager, database etc. The hub behavior is defined through a set of rules implemented in a special scripting language.

A speech-to-speech translation system can be seen as a special case of an SDS, consisting of interacting modules, but without dialog manager or database backend [Gruhn 01b, Gruhn 01c]. As such systems are not necessarily applied in fixed places where computers with large processing power are available, but "in the field", access to translation servers from lightweight clients is a key feature. Popular approaches include cellular phones [Gruhn 00b, Gruhn 00a] or small-scale portable computers such as PDAs [Gruhn 99, Singer 99a].

User utterances are recognized, translated to a pre-selected target language and then played back with a TTS. To reduce the effect of recognition and translation errors, feedback such as display of the recognized result is crucial. SMILE is a translation system based on the architecture standard CORBA (Common Object Request Broker Architecture [OMG 90]), supporting various input types, including close-talking microphone and telephony hardware. As shown in Fig. 2.4, the modules are operating independently and event-driven without a managing module to steer the flow of information. The client interface is a reporting tool that provides final and intermediate results to the conversation partners. Each of the modules can be included several times on distributed servers for high-speed performance.

While evaluating a speech recognition system with correctness and accuracy is quite straightforward, evaluating a speech dialog system [Minker 98a, Minker 02a] or a speech-to-speech translation system [Sugaya 00] is still a open research topic.

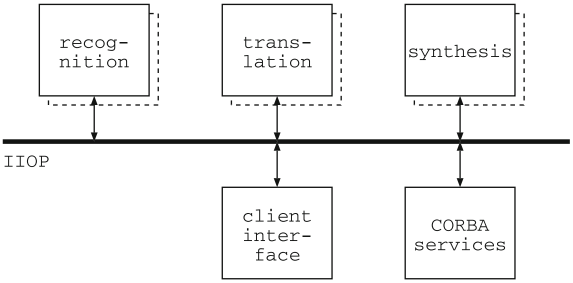


Fig. 2.4 The speech and multimodal interface for multi-lingual exchange (SMILE) architecture

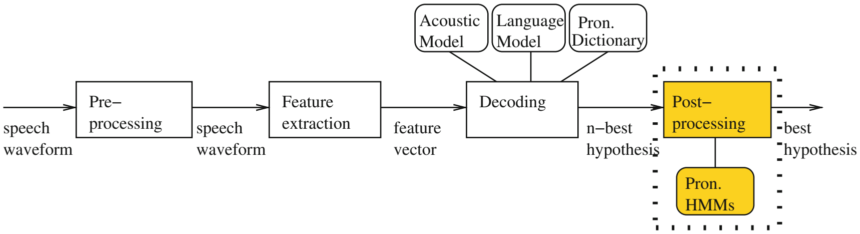


Fig. 2.5 The focus in this work lies on post-processing

Additionally to the speech recognition evaluation, task completion rate is a popular measure: The share of dialogs in which the human customer was able to achieve his goal, such as getting a specific required information or booking a hotel room.

2.5.3 Focus of This Work

This research bases on a standard automatic speech recognition system. The modules from speech recording to decoding follow the common practices as described in this chapter. Our novel approach described in Chap. 7 focusses on the post-processing part, as shown in Fig. 2.5. Here we apply discrete HMMs as statistical pronunciation models for rescoring an n-best list. These discrete models are trained as described in Sect. 2.1.3 with the result of a phoneme recognition as input data.

Chapter 3
Properties of Non-native Speech

A typical example for a speech recognition system that needs to understand non-native speech is a car navigation system. In Europe, a driver who travels across borders and inputs a destination city or street name of a foreign country into a speech-controlled navigation system, is likely to mispronounce. In the United States, the high number of immigrants with limited English skills, and foreign-motivated place names are similar issues. A common workaround is to have users spell the names, but this is a less comfortable way than just saying the place name, and sometimes the correct spelling of a place name is not known either.

Non-native pronunciations are also a big issue for speech dialog systems that target tourists, such as travel assistance or hotel reservation systems. Usually, they provide only a few languages, or only the local language and English, and travelers from other countries are expected to get the information they need with the English version. While many travelers do speak English to some extend, the non-native speech poses several challenges for automatic speech recognition.

3.1 Accent

People from different countries speak English with different accents, depending mainly on the mother tongue and the foreign language [van Compernolle 01]. Other factors such as skill also have some effect. A person can understand non-native speech easily because after a while the listener gets used to the style of the talker, i.e. the insertions, deletions and substitutions of phonemes or wrong grammar. The vocabulary and grammar of non-native speakers is often limited and easy, but a recognizer takes no or only little advantage of this and is confused by the different phonetics.

In literature, there are several terms to describe the language pair elements involved in non-native speech. The mother tongue of the speaker is called main

R.E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*, Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_3,

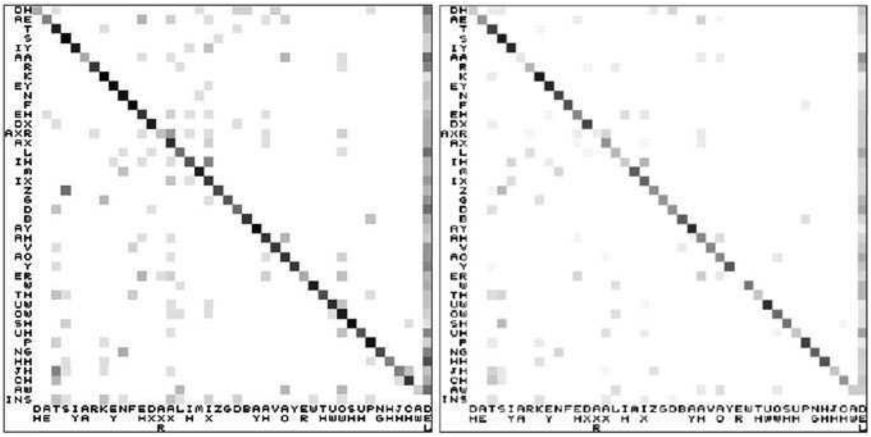


Fig. 3.1 Phoneme confusion matrix for German (*left*) and Japanese (*right*) speakers of English. The x-axis shows the recognition result, the y-axis the correct phoneme

language, source language or L1. The language the speaker attempts to speak but has not fully mastered yet is called the target language or L2. In papers focusing on computer aided language learning, the terms L1 and L2 are more common.

For computer assisted language learning systems, minimal word pairs are of special relevance. These are word pairs, that differ only in one sound, and for each accent specific word pairs can be found where the discriminating sound is especially difficult to produce for people with specific accents. A well known example is the discrimination between /L/ and /R/ in Japanese, one minimal word pair for this accent would be *read* and *lead*.

To find out, which sounds are frequently confused for a given L1 - L2 pair, it can be helpful to look at phoneme confusion matrices.

3.2 Phoneme Confusion

Figure 3.1 shows a phoneme confusion matrix comparing a phoneme recognition result with a reference phoneme transcription. The notation is in ARPAbet symbols (cf. Sect. 2.2). The darker a box is, the more frequently the regarding confusion occurred. The phoneme recognition result represents which sounds the speaker actually produced. The diagonal shows all cases of correct pronunciation, all other entries are mispronunciations. This simple model is of course somewhat distorted by random recognition errors, and all graphs show standard confusions like between the very similar sounds /m/ and /n/. Still the graphs show clearly some typical error patterns that are accent specific.

German speakers, for example, have difficulties producing a distinct /Z/ sound, it sounds similar to the voiceless /S/. This confusion does also occur for Japanese

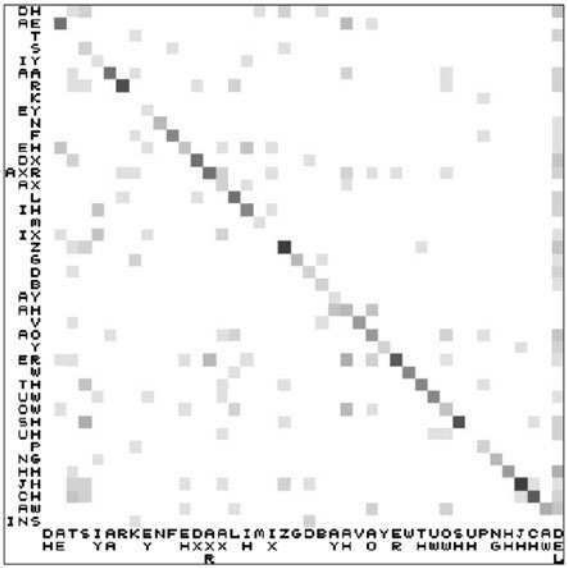


Fig. 3.2 Difference matrix comparing native speakers of English with Japanese accented English. The darker, the greater the difference

speakers, but less frequently so. The reason is that the German language does not have a /Z/ sound, whereas Japanese has both /S/ and /Z/. Japanese speakers of English have problems producing the /R/ sound, the phoneme recognizer classifies it as /L/ or a wide range of other phonemes. Again the reason is that the Japanese language has no /R/. While the graphs also contain some random confusions, it is still visible that there are some accent-specific confusion patterns.

The confusion matrices have been calculated on the ATR non-native speech database, which contains also three native speakers of English, one Australian, one British and one US-American. The data of three speakers of rather different types of English may not be the ideal basis to calculate a confusion matrix from, but for the purpose of illustration we calculated a phoneme confusion matrix for native English as well. Subtracting the confusion values of the native speakers from one of the accents' confusion matrix and calculating the absolute values matix yields an image visualizing the differences between native and non-native pronunciation and recognition properties. An example for such an image is Fig. 3.2, where we compare Japanese English to a native English matrix. Appendix B shows more examples for phoneme confusion matrices..

3.3 Non-phonemic Differences

Non-native speakers have a limited vocabulary and a less than complete knowledge of the grammatical structures of the target language. The limited vocabulary forces speakers to express themselves in basic words, making their speech unusual for native speakers. At the same time this effect can make non-native speech easier to understand for other non-natives.

Vocabulary limitation tends to be a less relevant issue for speech dialog systems. In many applications the vocabulary is fixed and known to the speaker. For example a user trying to input a destination in a foreign country may not know the correct pronunciation of the place name, but he will know the orthography. A similar effect occurs for a speech controlled media player: Even if the user has difficulties translating a song title in a foreign language correctly, the word sequence is usually known if the user has copied it to his media player.

The same holds for grammatical problems, for an address there is no grammar and the word sequence of a song title is fixed and does not require the user to have mastered the target language to reproduce.

For a travel arrangement task, both vocabulary limitation and grammar can be issues. As long as the relevant keywords can be assumed to be known, keyword spotting based algorithms are capable of circumventing grammatical problems. If the vocabulary of the speaker does not allow to complete the task, providing alternative language model options might help. But it will be difficult to cover all possible variations how to describe unknown vocabulary.

A different major challenge about non-native speech processing lies in prosodic effects and other non-phonetic differences like:

- lengthening
- hesitation
- stops
- filled pause insertion

As they occur randomly at any position in an utterance they can be hard to predict and handle.

Word or phoneme lengthening is to some extend covered by the HMM architecture, as an increased self-loop time has little impact on the acoustic score. Fillword insertions and hesitations can be permitted by a language model at any position, but too long hesitations can mislead an endpoint detection module into declaring an end of the utterance. The initial part of the utterance alone is less likely to be meaningful and a misrecognition may occur, if the user continues his utterance later, the second half may be misrecognized as well. Configuring endpointing to be lenient about long pauses can address this problem, but if background noise is an issue as well, an overall decreased recognition accuracy is likely. If a user stops in the middle of a word, for example to review its pronunciation or to think about how to continue an utterance, the regarding word will probably be misrecognized.

The more spontaneously the utterance is spoken, the more likely such nonphonetical problems are to occur. Read speech, or speech control of media and navigation devices, is less likely to encounter those effects as the user need to think less about what to say. But even with the relative absence of fillwords and hesitations in such a more limited scenario, the pronunciation variations in non-native speech are a big challenge.

In this book, we concentrate on the pronunciation issues of non-native speech. In order to separate spontaneous speech issues from pronunciation, we work on read speech. This is a realistic scenario related to real-world applications like media control or navigation. It is also the part where speech recognition technology as possibilities to recover from errors, whereas helping the user finding the right words to say or to produce his utterance in a controlled matter is more a design issue for dialog system construction.

Chapter 4
Pronunciation Variation Modeling in the Literature

In an automatic speech recognition system, speech properties are represented usually by an acoustic model and a pronunciation dictionary. But every human has a slightly different way to speak words. Those differences must be incorporated in the underlying speech models.

While this problem is extreme for non-native speakers, it is also an issue for native speakers, due to regional dialects that create variations in pronunciation, or even pronunciation patterns that depend only on the individual. An example in the German language is the /g/ e.g. at the end of the word *richtig*, which can be pronounced either as [g] or as [ç] (in IPA-notation), depending only on the speakers personal preference. Much literature is therefore analyzing pronunciation variations in native speech.

On the acoustic model level, the usual approach is to collect a very large speech database. If typical variations are sufficiently covered in the training data of the acoustic model, speaker variations are included in the model statistics. Although there is no clear definition of when a database is sufficient, experience shows that acoustic models trained on large databases are quite reliable.

To target the properties of specific individual speakers, acoustic model adaptation techniques are available. The most common are *maximum a-posteriori* (MAP) and *Maximum Likelihood Linear Regression* (MLLR). In both cases, the user is required to utter some training sentences. The Gaussian density parameters of a speaker independent acoustic model are adapted on this small data set. The result is an acoustic model that is especially suitable to cover the special properties of that speaker, but has lost general applicability.

Pronunciation lexicon adaptation methods can be grouped based on three key properties [Strik 99]:

- The types of variations covered
- The methods to obtain the pronunciation information
- The representation of the information

R.E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*,
Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_4,

4.1 Types of Pronunciation Variations

Pronunciation variations can include phoneme substitutions, deletions and insertions. These phoneme-level variations can be applied to given pronunciations in a dictionary, adding additional entries. Most research follows this approach, because it is easy to implement.

Cross-word variations also occur, especially in fast spontaneous speech. They can be modeled by adding multi-word entries to the dictionary [Blackburn 95]. Of course only the most frequent expressions can receive this special treatment, otherwise the dictionary size would increase too much.

4.2 Data-Driven and Knowledge-Based Methods

Knowledge about which pronunciation variations occur can be found in linguistic literature or analysis of manually produced pronunciation dictionaries. It can also be extracted in a data-driven way.

The data-driven method requires a speech database, ideally with phonetic labels. If no such labeling is available, it can be generated automatically. A speech recognizer with a list of phonemes as dictionary produces a sequence of phonemes. Usually a phoneme bigram provides some phonotactic constraint. Such automatically generated phoneme sequences always contain recognition errors, but as large databases can be transcribed this way, and random errors can be averaged out. Also, as manual phonetic transcriptions are subject to the individual perception of the human labeler, manual transcriptions are not free of error either [Cucchiarini 93]. Errors that occur frequently and systematically are the very information desired. Alternatively to phoneme recognition, a phonetic transcription can also be achieved with forced alignment [Riley 99].

To apply the recognition results, the most frequent phoneme sequences of each word can be collected to create a new pronunciation dictionary or to enhance an existing one. The most frequent application is to compare the phoneme recognition result to canonical pronunciations in a baseline dictionary by dynamic programming. The resulting alignment serves as basis to generate the desired representation of pronunciation variations, most frequently phoneme confusion rules.

Comparing data-driven and knowledge-based approaches [Wester 00, Witt 99b], both can lead to roughly the same information. Therefore, data-driven approaches are more desirable, as they require less human expert knowledge and can easily be applied to many accent types.

In addition to provide information about pronunciation variation effects, data-driven approaches are also affected by properties of the specific speech recognition system (especially the acoustic model). Single insufficiently trained HMMs in an acoustic model can produce phoneme confusion patterns, which are difficult to distinguish from actual pronunciation variations. Compensating for these errors in the dictionary leads to additional improvement in recognition accuracy.

4.3 Representing Pronunciation Variations

4.3.1 Phoneme Confusion Rules

The most common approach to represent pronunciation variations is in the form of rules, which are applied by generating additional entries in the pronunciation dictionary [Binder 02, Sloboda 96, Yang 00, Amdal 00a]. These pronunciation rewriting rules r can be formalized as [Cremelie 97]:

$$r : LFR \rightarrow F' \text{ with } P_r \tag{4.1}$$

with F, L, R and F' being variable length phoneme sequences, also called pattern. P_r is the firing or activation probability of the rule. A phoneme sequence F with left context L and right context R is transformed into a sequence F'. The firing probability can be calculated in data-driven approaches and are a measure of how likely a pronunciation variation occurs. Based on these firing probabilities, each additional dictionary entry can be assigned a pronunciation probability, which is helpful to scale the number of pronunciation variations added to the dictionary.

Many dictionaries already include some common pronunciation variants, such as the two pronunciations /i:D@r/ or /aID@r/ for the word `either`.

Experiments have shown that adding pronunciation variations to a dictionary does have a big impact on the performance. But there is also a maximum average number of pronunciations per word up to which recognition accuracy is increasing. If too many pronunciations are added, additional confusions occur and recognition accuracy deteriorates. Choosing the right pronunciation probability threshold(s) is therefore a very important but tricky issue. Depending on the experimental setup, this number lies between 1.1 and 1.5. There are several way to apply firing probabilities of rules to scale the dictionary:

- Ignoring all rules with a firing probability below a certain threshold [Binder 02]
- Likewise, but setting separate thresholds for deletion, substitution and insertion rules [Yang 02]
- Including only additional pronunciations of words for which the overall pronunciation probability is sufficiently high.

An additional measure is the frequency of words in a large text corpus, words that rarely occur at all should not cause too many entries in the dictionary. The confusability between variants can also serve as criterion [Sloboda 96].

For rather homogenous speech such as the native speakers of the Wall Street Journal database, adding pronunciation variants to the dictionary can show some improvement but it is likely to be small [Yang 02].

For more speaker groups with stronger pronunciation deviations, such as non-native speakers, the potential to improve recognition accuracy with this approach is higher [van Compernolle 01, Goronzy 01].

4.3.2 Pronunciation Networks

One possibility to generate pronunciation variations for a recognizer dictionary is with neural networks. As with rules, output from a phoneme recognizer and canonical transcription are aligned and the correspondence between the two pronunciations used for neural network training. For each phoneme in a quinphone context, there is an input unit. On the output layer, there is an insertion, deletion and substitution unit for each phoneme. Applying the network on the baseline dictionary leads to some word accuracy improvement [Fukada 97].

As HMMs have shown to be reliable for acoustic modeling, examining their applicability to model pronunciation variations seems promising.

A simple idea is to build an acoustic model with words rather than phonemes as base units to deal with multiple non-native accents. If the vocabulary size in a recognition task is very small, such modeling is feasible and helpful [Teixeira 97], but for large vocabularies, it will be difficult to find enough training data to reliably train acoustic word models.

Therefore it seems more promising to keep acoustic modeling and lexicon construction separate, e.g. with HMMs for lexicon construction.

The concept of generating HMMs to model pronunciation has been analyzed for automatically generated acoustic subword units. This method has been applied to an isolated word task with one Norwegian speaker [Paliwal 90] to generate pronunciation dictionaries and for a database of 150 Korean speakers [Yun 99].

4.4 Other Related Topics

4.4.1 Computer Assisted Language Learning (CALL)

Learning a language requires expert knowledge to learn from, which can come from a human teacher or a textbook. But human teachers are not always available and costly, while textbooks alone can not provide feedback on pronunciation. A computer program could provide all of these desired features, if a method to provide feedback is found. Such software would make language study at home much more effective. Research on computer aided language learning systems therefore focuses on methods to automatically analyze pronunciation skill.

CALL systems ask speakers to utter text displayed on the screen and analyze it. The text can be whole sentences, giving the learner feedback on his overall pronunciation. To evaluate specific pronunciations, it is common to examine words or minimal word pairs (i.e. two words that differ only in a single phone, such as *lead* and *read* [Akahane-Yamada 04].

Many CALL systems apply a native HMM acoustic model, perform a forced alignment on the utterance and calculate a rating of the speakers skill based on the acoustic score, which is a probabilistic measure of similaristy to the native speech

[Bernstein 90]. Additional information can also be extracted, e.g. speaking rate and pause have been shown to be interesting features [Tomokiyo 00, Kato 04].

The quality of the automatic skill *rating* is evaluated by comparing it to *gradings* by human experts (such as language teachers). Unfortunately, human graders agree on speakers skills only to some limited extend, inter-grader correlations tend to be around 65 %[Neumeyer 00, Cincarek 04a]. The correlation between ratings and gradings is accordingly somewhat lower than this value. It is therefore difficult to prove the validity of automatic ratings.

A different approach to measure nativeness is to recognize with a network of two acoustic models in parallel, either native and non-native, or first and second-language. A general rating can be calculated on which models the decoder applied more often. The feedback of such an approach is especially valuable, as the system can then point out to the user which sounds require additional attention [Kawaii 99]. Minimal word pairs are of high importance for this task.

4.4.2 Multilingual Speech Recognition

The topics non-native speech recognition and multilingual speech recognition are sometimes confused. Multilingual speech recognition is any technology that involves more than one language; if only two languages are considered the research can be called bilingual.

Approaches typically focus on the acoustic model, as speech data is expensive to collect and a method to generate an acoustic model without training would be desirable. Such a model assembled from other languages is called cross-lingual.

After the discovery that the acoustic similarity of phonemes across language can be exploited for building cross-lingual acoustic models [Köhler 96, Gruhn 01a], much research was conducted to create acoustic models of languages, for which no or little speech data is available [Schultz 00b, Schultz 01]. Such rapid prototyping is especially helpful to create ASR systems for minority languages or languages of third-world countries where funds for data collection are not available; recognition rates are usually lower than those of ASR systems trained on a large speech database. As phonetic contexts differ very much across languages, such research tends to base on monophone models The users of such ASR systems are assumed to be native speakers.

The typical approach to generate a multilingual acoustic model bases on a list of phonemes in the source language(s) and the target language. For each phoneme of the target language, a phoneme from one of the source languages is chosen. The choice is often based on matching symbols in a phonetic alphabet. However, if some data of the target language is available, similarity can also be determined in a data-driven way. The acoustic models for the phonemes are combined together into one model set, which is suitable for recognition in the target language to some extend. In order to achieve matching acoustic conditions, a multilingual database such as Globalphone [Schultz 00a] or CALLHOME [Billa 97] with similar or identical recording setup for each language is advantageous, somewhat tainting the aim to avoid costly data collections.

Non-native speech can be considered a special case of this approach, an own language for which little data is available. Acoustic models for non-native speech can be generated by combining the acoustic models of native and foreign language. Those bilingual models are accent-specific, but can capture the special properties of that accent quite well [Übler 98.]. Most multilingual research seeks to create an acoustic model for native speech without training data for that language, which succeeds but leads to recognition rates lower than what can be expected from a properly trained model. But for non-natives, accuracies higher than what can be achieved with native models have been reported with multilingual models [Übler 01].

Research has also been conducted in interpolating acoustic models of native and foreign language [Steidl 04], leading to good results for specific accents. The approach of generating a non-native pronunciation dictionary by recognizing native speech of the target language with a phoneme recognizer of a source language has also been analyzed, with relative word improvements of 5.2% relative [Goronzy 04].

There have been several attempts to make general purpose acoustic models more suitable for recognition of non-native speech with the model adaptation techniques MLLR or MAP, with a native speech database of the target language as adaptation material. The advantage of such an approach is that there is no need to collect a non-native database for training. The disadvantage is that the adaptation cannot take into account which errors non-native speakers make and which sounds are hard to pronounce. Literature is somewhat inconclusive about whether such an approach can be effective or not, with one paper reporting amazing improvements [Mengistu 08] and others being much less enthusiastic [Tomokiyo 01, Schultz 01].

Our approach of multilingual weighted codebooks described in Sect. 6.2 also targets to exploit multilingual similarity for non-native speech recognition.

4.5 Relevance for This Work

For our work, we want to cover all types of pronunciation variations. While expert knowledge might bring more accurate information about a given specific accent, it does not generalize, making data-driven analysis the preferred way of obtaining information. As form of representation, we will examine rules which we apply for lattice processing in Sect. 6.1. However, because of the limitations of rules for representing non-native variations, we propose a pronunciation network style approach with HMMs as statistical lexicon in Chap. 7.

We will examine a multilingual approach for non-native speech recogntion in Sect. 6.2 with multilingual weighted codebooks for semi-continuous HMMs. Features to provide a skill scoring of non-native speakers as well as detecting mispronounced words automatically are examined in Sect. 6.3.

Chapter 5
Non-native Speech Database

One of the contributions of this work to the research community is the design, recording and skill annotation of a large database of non-native English speech uttered by speakers from several accent groups. This chapter gives an overview of this database. We also summarize existing non-native speech databases and a provide classification into database types, explain their limitations and the importance of collecting our own database. All important details about data recording setup and properties of the *ATR non-native database* collected in this work are described in this chapter, some additional information of interest can be found in the appendix.

5.1 Existing Non-native Speech Databases

Although the number of existing databases specializing on non-native speech is considerably lower than the number of standard speech databases, several such special speech data collections already exist.

It does not seem uncommon that a paper presenting results on non-native speech is based on a private data collection of the regarding authors or institute. So far, no "standard" database for non-native speech research has been generally acknowledged in the research community, unlike for example TIMIT [Fisher 87] for phoneme recognition or the AURORA series [Hirsch 00] for digit recognition and noise compensation research. This has very serious implications:

- It seems difficult to compare different approaches towards non-native speech recognition and judge the validity of proposed new methods.
- Collecting a new database is expensive (in terms of time and money resources). This leads to much research being published on tiny databases or even indi-

R. E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*, Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_5,

vidual speakers or data that is recorded under questionable conditions. The reliability of results calculated on too small databases is doubtful.

There have already been serious efforts to collect non-native speech databases, which we discussed in [Raab 07].

Most of the corpora are collections of read speech, and the utterances are collected from speakers of non-native English.

Table 5.1 gives detailed information about each corpus: The name of the corpus, the institution where the corpus can be obtained, or at least further information should be available, the language which was actually spoken by the speakers, the number of speakers, the native language of the speakers, the total amount of non-native utterances the corpus contains, the duration in hours of the non-native part, the date of the first public reference to this corpus, some free text highlighting special aspects of this database and a reference to another publication. The reference in the last field is in most cases to the paper which is especially devoted to describe this corpus by the original collectors. In some cases it was not possible to identify such a paper. In these cases the paper where we first found information about this corpus is referenced.

Some entries are left blank and others are marked with unknown. The difference here is that blank entries refer to attributes we were not able to find information about. Unknown entries, however, indicate that no information about this attribute is available in the database itself. Example the MIT Jupiter corpus [Livescu 99] has been collected as an open public system providing weather information to random users, leading to a certain share of non-native speech. Due to the nature of collection, there is no information available about the speaker properties other than a non-nativeness flag set by the transcribers. Therefore this data may be less useful for many applications. In the case where the databases contain native and non-native speech, the we aimed to only list attributes of the non-native part of the corpus.

Onomastica [Onomastica 95] is not a speech database, but a non-native pronunciation dictionary. We included it in this table, because due to its large size it is a significant resource for non-native speech and deserves a reference.

Where possible, the name is a standard name of the corpus, for some smaller corpora, however, there was no established name and hence an identifier had to be created. In such cases, we chose a combination of the institution and the collector of the database.

In order to provide the research community with an always up-to-date reference list about non-native speech databases, a website was created at the English language Wikipedia [Wikipedia 07] website where interested researchers can browse for databases that fit their purposes or add information about new collections. Keeping track permanently of every data collection in the world is hardly possible for a single person. Therefore Wikipedia was chosen as publication method so that anybody who detects missing information can directly edit and update the table without permission problems.

Table 5.1 Overview of non-native Databases. Language abreviations are A:Arabic, C:Chinese, Cz:Czech, D:Danish, Du:Dutch, E:English, F:French, G:German, Gr:Greek, H:Hebrew, In:Indonesian, I:Italian, J:Japanese, K:Korean, M:Malaysian, N:Norwegian, P:Portugese, R:Russian, S:Spanish, Sw:Swedish, T:Thai and V:Vietnamese

Corpus	Author	Available at	Language(s)	Speakers	Native language	#Utt.	Hours (h)	Date	Specials	Reference
ATR-Gruhn	Gruhn	ATR	E	96	C G F J In	15000	22	2004	Proficiency rating	[Gruhn 04a]
BAS Strange I+II		ELRA	G	139	50 countries	7500		1998		[Munich 98]
Broadcast News		LDC	E					1997		[Tomokiyo 01]
Berkeley Restaurant		ICSI	E	55	G I H C F S J	2500		1994		[Jurafsky 94]
Cambridge-Witt	Witt	U. Cambridge	E	10	J I K S	1200		1999		[Witt 99a]
Cambridge-Ye	Ye	U. Cambridge	E	20	C	1600		2005		[Ye 05]
Children News	Tomokiyo	CMU	E	62	J C	7500		2000	Partly spontaneous	[Tomokiyo 01]
CLIPS-IMAG	Tan	CLIPS-IMAG	F	15	C V		6	2006		[Tan 06]
CLSU		LDC	E		22 countries	5000		2007	Phone, spontaneous	[Lander 07]
CMU		CMU	E	64	G	452	0.9		Not available	[Wang 03]
Cross Towns	Schaden	U. Bochum	E F G I Cz Du	161	E F G I S	72000	133	2006	City names	[Schaden 06]
Duke-Arslan	Arslan	Duke Univ.	E	93	15 countries	2200		1995	Partly phone speech	[Arslan 97]
ERJ	Minematsu	U. Tokyo	E	200	J	68000		2002	Proficiency rating	[Minematsu 04]
Fitt	Fitt	U. Edinburgh	F I N Gr	10	E	700		1995	City names	[Fitt 95]
Fraenki		U. Erlangen	E	19	G	2148				[Stemmer 01]
Hispanic	Byrne		E	22	S		20	1998	Partly spontaneous	[Byrne 98]

(continued)

Table 5.1 (continued)

Corpus	Author	Available at	Language(s)	Speakers	Native language	#Utt.	Hours (h)	Date	Specials	Reference
IBM-Fischer		IBM	E	40	S F G I	2000		2002	Digits	[Fischer 03]
ISLE	Atwell	EU/ELDA	E	46	G I	4000	18	2000		[Menzel 00]
Jupiter	Zue	MIT	E	unknown	unknown	5146		1999	Telephone speech	[Livescu 99]
K-SEC	Rhee	SiTEC	E	unknown	K			2004		[Rhee 04]
MIST		ELRA	E F G	75	Du	2200		1996		[Institute 07]
NATO HIWIRE		NATO	E	81	F Gr I S	8100		2007	Clean speech	[segura 07]
NATO M-ATC	Pigeon	NATO	E	622	F G I S	9833	17	2007	Heavy noise	[Pigeon 07]
NATO N4		NATO	E	115	unknown		7.5	2006	Heavy noise	[Benarousse 99]
Onomastica			D Du E F G Gr I N P S Sw			(121000)		1995	**Only lexicon**	[Onomastica 95]
PF-STAR		U. Erlangen	E	57	G	4627	3.4	2005	Children speech	[Hacker 07]
Sunstar		EU	E	100	G S I P D	40000		1992	Parliament speech	[Teixeira 97]
TC-STAR	Heuvel	ELDA	E S	unknown	EU countries		13	2006	Multiple data sets	[van den Heuvel 06]
TED	Lamel	ELDA	E	40(188)	many		10	1994	Eurospeech 93	[Lamel 94]
TLTS		DARPA	A		E		1	2004		[Mote 04]
Tokyo-Kikuko		U. Tokyo	J	140	10 countries	35000		2004	Proficiency rating	[Nishina 04]
Verbmobil		U. Munich	E	44	G		1.5	1994	Very spontaneous	[Munich 00]
VODIS		EU	F G	178	F G	2500		1998	Car navigation	[Trancoso 99]
WP Arabic	Rocca	LDC	A	35	E	800	1	2002		[LaRocca 02]
WP Russian	Rocca	LDC	R	26	E	2500	2	2003		[LaRocca 03]
WP Spanish	Morgan	LDC	S		E			2006		[Morgan 06]
WSJ Spoke			E	10	unknown	800		1993		[Amdal 00b]

Non-native speech research can be divided into several typical groups. It plays a role in the research areas of speech recognizers, text to speech systems, pronunciation trainers or computer assisted language learning systems. The task might be to train, to adapt or only to test a system.

The best way to classify non-native speech databases is regarding to the type of application they are designed for. The major fields are navigation devices or travel assistance, military communications, presentation systems or computer assisted language learning systems.

5.1.1 Speech Operated Travel Assistance

A possible future application of non-native speech recognizers are automatic tourist information or hotel booking systems. As they are unlikely to cover any language in the world, to interact with the system many travelers will have to speak in English—non-native English.

A present application are navigation devices, where users can input destinations in foreign countries by speech. As most mobile devices they still have to cope with limited computing power. Therefore systems running on these devices are less elaborated and do not allow natural spontaneous input.

Of course, of major interest for both systems are city and street names as well as digits, for example for street numbers or postal addresses. Hence, a very interesting corpus for this task would be the CrossTowns corpus, as it covers mainly city names in a couple of languages. The strength of this corpus is that it includes many language directions (speakers of one native language speaking another language). Altogether the corpus covers 24 different language directions. Each recording of a language direction contains two times 45 city names per speaker. First the 45 city names are read from a prompt, and then they are repeated after listening to the name via headphone. 13000 of the utterances are manually transcribed at the phonetic level, and there is information about the language proficiency of the speakers. A planned release at the European language resource agency (ELRA/ELDA) in 2006 did not succeed. Unfortunately, licensing issues prevent its public release, so that researchers outside the originating university cannot access it. According to the author of the corpus a future release of this corpus is undetermined.

Two further corpora exist for this domain: CLIPS-IMAG and the ISLE corpus. The CLIPS-IMAG corpus is a collection of read speech with the advantage of covering the tourist domain, which is likely to contain similar places of interest as they will be demanded from navigation devices. With a total amount of 6 h of non-native speech this corpus is also relatively large. The disadvantage is that it consists of exotic language pairs (Chinese and Vietnamese speakers of French) that may be of limited interest for most researchers or commercial developers.

Compared to the other databases applicable for this field of research, the ISLE corpus has the disadvantage not to contain in-domain data, as it is more designed for CALL research (see below). Yet about half of the corpus are simple and short utterances, which is at least somewhat comparable to plain command interaction simple navigation systems can handle. ISLE is one of the largest non-native speech corpora and has the advantage to be distributed by ELDA for a moderate price. There are only two accents, German and Italian accented English in this corpus. The speakers read 1300 words of a non-fictional, autobiographic text and 1100 words in short utterances which were designed to cover typical pronunciation errors of language learners. The corpus is annotated at the word and at the phone level, which makes it especially interesting for the development of Computer Assisted Language Learning systems.

Because of the wide range of applications for call centers and automatic call processing systems it is unlikely to find a non-native speech database that precisely matches a given task, but from the viewpoint of domain, the travel assistance databases are probably the closest match. Considering acoustic properties, the non-native speech part of the Jupiter [Livescu 99] corpus is the largest fitting database, although it still contains comparatively few utterances and lacks information about the type of accent.

5.1.2 Military Communications

For automatic documentation of communication between soldiers from different member countries and other purposes, the military has a strong interest in speech recognition systems that are capable of handling non-native speech. To encourage research in this field, NATO has recently released a couple of interesting corpora. The M-ATC (Military Air Traffic Control) [Pigeon 07] covers pilot controller communications with a variety of accents, strong background noise and a high number of different speakers. The N4 corpus [Benarousse 99] contains recordings from naval communication training sessions in the Netherlands. The transcriptions of the N4 corpus are very rich regarding information about speaker background. The airplane engine and communication channel noise in the N4 and M-ATC corpora is so strong that several utterances were too unintelligible to be transcribed by human listeners. This makes the corpus a very tough task and more relevant for work about noisy speech, for which in turn the SPINE corpus [Zhang 02, Markov 03] of *speech in noisy e*nvironment might be more interesting.

The Hiwire corpus contains spoken pilot orders that are input for the Controller Pilot Data Link Communications [FAA 07]. An advantage of this corpus compared to the two previously mentioned ones is that the recordings were originally made in a studio. Thus this corpus provides clean speech as well noisy speech which was obtained through convolution of clean speech and noise. The Hiwire and M-ATC corpus yield the additional advantage to be free of charge for European researchers. Besides the noise, the military corpora have the disadvantage of containing very particular vocabulary, such as military keywords and geo positions.

5.1.3 Presentation Transcription

There are two databases that are likely to be useful for this application, namely TC-STAR [van den Heuvel 06] and TED [Lamel 94]. The TC-STAR corpus contains about 100 h of Spanish and English transcribed parliament speech each. It is not a specific non-native database, but it includes 11 h of non-native English and some amount of non-native Spanish in both training and test corpora of TC-STAR. A larger part of the TC-STAR corpus is from non-native interpreters. As it is not clear to what extent speech from a professional interpreter can really be considered non-native speech the non-native interpreter part is not included in this number. The speech is very spontaneous, containing restarts, hesitations and reactions to the audience.

The Translanguage English Database (TED) is a corpus which contains almost all presentations from the Eurospeech 1993. The speech material totals 47 h, however only about 10 h are transcribed. Due to the typical mixture of presentations from a variety of countries, it is believed that a large amount of the presentations is given with non-native accents. The speech contains all artifacts found in conference presentations, such as nervousness, room acoustic and spontaneousness effects.

Databases are frequently not collected in order to achieve optimal data quality that allows clear conclusions which method to deal with non-native effects. Rather, databases are created where they are cheap to collect. Transcriptions from meetings or political speeches are typical examples: The spontaneous speech effects and acoustic condition as well as the unclear accent status cause these databases to be rarely employed for non-native speech research.

5.1.4 Computer Assisted Language Learning

Most speech technologies only need orthographic transcriptions for the databases to train systems. This is different for computer aided language learning (CALL) systems. In order to detect and/or classify mispronunciation it is necessary to have ratings of overall skill and pronunciation quality, as well as some kind of information about pronunciation correctness at word level, either in form of markings of incorrect words or in form of a transcription at the phonetic level. The quality of these ratings is of utmost importance, as a typical target of CALL research is to create an automatic scoring algorithm that matches the human ratings as closely as possible, therefore resembling the opinion of human experts. Major corpora which can provide proficiency ratings are ISLE, Cross Towns, and ERJ. Of these corpora, the ISLE and Cross Towns corpus contain also transcriptions at the phonetic level. The database of English read by Japanese (ERJ) has been collected at the University of Tokyo with the specific target of aiding CALL research. It focuses on Japanese accented English. ISLE, as described in Sect. 5.1.1, was designed for the purpose of CALL research, making it its main field of application. Cross Towns would also be a good corpus if it were publicly available.

5.2 The ATR Non-native Speech Database

A significant part of the work presented herein is the collection of a non-native speech database. The collection was motivated by the problem that much research on non-native speech is published with results calculated on individual databases that in many cases are too small or unreliably collected, making it very hard to judge the effectiveness of the proposed methods. Another important point is to cover several different accent types. Many approaches need much expert knowledge about the properties of a specific accent, and do not generalize as would be needed for a successful commercial system dealing with non-native speech in many languages. Furthermore, the database should be as versatile as possible, allowing for non-native speech recognition based applications as well as CALL research. For the latter, ratings by human experts are required. Finally, the database must be made available to the research community in order to avoid creating yet another local database.

5.2.1 Speakers

This ATR (Advanced Telecommunication Research Laboratory) non-native speech database consists of 96 non-native speakers of English, covering many types of non-native accents. Most speakers come from China, France, Germany, Indonesia or Japan, with the Japanese speaker group being slightly larger than the others, as the data was collected in Japan. The mother tongue of all subjects listed under Chinese was Mandarin Chinese. Seven speakers have other native languages: Bulgarian, Spanish, Portuguese, Hindi, Sinhalese, Hungarian or Korean. The age range is 21–52 years, but most speakers are around 30 years or younger. Table 5.2 lists those speakers, further information about the individual speakers can be found in Appendix C. Additionally, there is data from three male native English speakers (one British, one Australian and one US-American citizen), but as they are not non-native, they are not counted for the database size.

About half of the subjects are or were members of ATR at the time of the recording. The remaining subjects in the database were hired through an agency. There was no specific criteria about how the speakers were chosen, leading to a wide skill range. One subject turned out to be hardly able to read Latin characters and was rejected. Other than that, data from all speakers who were able to utter the complete text set within 2 h was included (a native speaker requires less than 30 min for this task). For a few speakers with extraordinarily low skill or who had to leave the session early, only a fragment is could be recorded. This data is not

Table 5.2 First language distribution and age ranges in the ATR non-native speech database

NNS-DB	# Male	# Female	Age
Chinese	16	2	21–52
French	15	1	21–42
German	15	0	23–43
Indonesian	15	1	24–43
Japanese	15	9	21–45
Other	5	2	31–42
Total	81	15	21–52

included in the official data set and was not employed for the work presented herein.

Our database is not gender-balanced. When the database collection plan was made, other experiments on non-native speech were also thought about, such as automatically clustering non-native speakers on the acoustic model level. To avoid that algorithm to only cluster by gender and to avoid the need to half the database for that experiment, we decided to focus on male speakers. The agency that provided the speakers found it easier to provide male subjects, too. Pronunciation variations are known to depend on factors like native language or learning experience, the gender has so far not been reported as a source of pronunciation problems. Hitherto we can assume it has no effect on our research to accept a gender imbalance.

5.2.2 Data Contents and Properties

Each subject reads a uniform set of around 150 sentences as listed in Table 5.3. It includes 25 credit card style number sequences, 48 phonetically compact sentences (a randomly chosen subset of the TIMIT SX set, see Sect. 7.5.1) and six hotel reservation dialogs with 73 utterances in total.

The C-STAR research consortium has been conducting research in speech-to-speech translation for travel assistance [C-Star 99]. In order to analyze such dialogs, they have collected and transcribed several conversations (between native speakers), giving one speaker the role of a traveler and the other speaker the role of a receptionist at a hotel. The transcriptions of five of these dialogs were included in the sentence list. Furthermore, one of the typical demonstration sample dialogs, “demo02”, was also added. A transcription sample from the hotel reservation dialogs can be found in Appendix A. Two of the hotel reservation dialogs, TAS22001 and TAS3202, were defined as test set of about three minutes, the rest of about eleven minutes as training data.

The sentence set was chosen based on which data is helpful for non-native speech research. These are some phonetically compact sentences for good phonetic coverage, and some sentences from the target scenario, which is hotel reservation.

Table 5.3 Detailed contents for each speaker of the ATR SLT non-native English database

Set	Contents	#Words	#Utterances
SX	48 phonetically rich sentences from the TIMIT database	395	48
TAC22012		252	19
TAS12008	Hotel reservation dialogs	104	9
TAS12010		144	12
TAS22001		162	10
TAS32002		182	13
demo02		70	10
DIGITS	Credit card numbers	200	25

One may argue that depending on the experience of the individual speaker some parts of the text set might be very difficult and the text set should take this into account. But before the recordings it is not clear to data collectors what tasks will be regarded as easy or difficult by the foreign speakers. Therefore we chose to pursue a uniform text set.

The recording equipment consisted of a Sennheiser HD-410 close-talking microphone and a Sony DAT recorder. The speech was recorded at 48 kHz and downsampled to 16 kHz with 16 bit precision. Of the hardware, only a display and mouse were in the vicinity of the speaker during recording. The PC was shielded from the recording booth to avoid background noises e.g. from the PC ventilation. To ensure no noise from the power network disturbs the recording hardware, the power supply was routed through a special power line network noise filter.

The recordings took place in two separate sessions. The first session consisted of eleven Japanese, two German and two native English speakers, the second session covered the rest. The recording setup is identical for the two sessions. For organisatory reasons, there was a time gap between the two sessions, therefore the smaller data set from the first session was already applied in some experiments (which are marked accordingly).

The recording software is interactive and controlled by the speaker. The software presents one sentence at a time to the speaker. Clipping and too low volume problems are automatically detected and indicated on a graphical user interface (GUI) to ensure acoustic quality. After uttering one sentence, the speaker could listen to the speech and record it again if he was not satisfied with the recorded utterance. Additionally, a recording supervisor asked the speaker to repeat if some noise or non-verbal human sound occurred during the speech, or the speaker made severe reading mistakes such as leaving out words.

If speakers felt unsure about how to pronounce a word, they were encouraged to attempt saying them the way they believe correct, as low confidence in the speakers' own language skill is a frequent problem. If the pronunciation of a word was completely unknown, a supervisor would assist.

5.2.3 Speaker Anxiety

A major issue in collecting non-native speech is to reduce the anxiety of the subject. Having to talk in a foreign language knowing that the speech is being recorded is a cause for stress. Less proficient speakers are more anxious about recordings of their speech as they regard the situation as a test of their proficiency of the foreign language. An extra factor may be that the speakers are being paid for the recording and feel pressure to produce good work. Unfortunately in a setup where the speakers are not paid it is difficult to accumulate a large number of motivated speakers, especially speakers other than students. Some initial recordings were conducted in an acoustic chamber. There, we observed that anxiety adds several additional artifacts to the speech:

- Sentences are artificially segmented into words.
- Speakers hyperarticulate words in an attempt to speak especially clearly.
- The speaking rate is—depending on the individual—much higher or lower than natural.
- Additional pronunciation mistakes occur because the speaker cannot remember the correct pronunciation.
- Due to lowered self-confidence, the speaker asks for help from the instructor more often.
- Nervous speakers also tend to get a sore throat quicker.

Such anxiety effects are a recording setup effect and are not necessarily expected in a real-world target system. In a speech-controlled car navigation for example, there is typically no critical audience, the user has as many attempts as he wishes and a haptical input system as backup, so there is no reason to be nervous. CALL systems are typically used alone, with no second person listening in front of whom the user could feel ashamed about his pronunciation. Presentation speech does contain anxiety, and because of the combination of non-native, spontaneous and anxious speech effects it is a very difficult task, discouraging the use of such databases at the current state of technology for non-native speech research. Military speakers such as flight controllers train the occurring situations, which can be expected to reduce nervousness. Also, they do not talk to a system but to a human dialog partner.

To reduce anxiety of the speakers, several measures were implemented. Although an acoustic chamber was available, we chose to record in a normal meeting room. An acoustic chamber is a room with walls especially designed to absorb sound and noise-free air supply. Because of the expensive wall padding necessary, acoustic chambers tend to be small. A person entering an acoustic chamber for the first time tends to feel uncomfortable or even scared because of the unnatural silence, an effect we had to avoid. Therefore we built a recording booth in a normal meeting room using sound-absorbing panels. The booth was half-open to the back and not fully closed on the top to avoid claustrophobic feelings. While significantly reducing reverberance, with this construction we could avoid the eerie atmosphere of a real acoustic chamber and were able to get both good data and comparatively relaxed speakers. The room had normal air conditioning. While we chose a location within the room with low noise, a minimum amount of background noise was unavoidable.

An instructor sat directly next to the speaker. The main role of the instructor was to give initial explanations how to use the recording equipment and software and to supervise the recording. However, a very important duty was also to reassure the speaker that he is doing all well and as expected. This can be achieved simply by nodding friendly whenever the speaker looks up. In case the instructor had to correct an utterance, special care was taken to be friendly and polite.

As a final measure, the subjects had to make pauses regularly and drink an alcohol-free beverage of their choice, which creates an opportunity to tell the speaker he is performing better than many other speakers, relaxes the speaker and counters sore throats.

5.2.4 Skill Annotation

As this database also aims to be applicable in CALL research, annotations of general skill and pronunciation errors are necessary. They are needed to validate or train an automatic scoring system.

Fifteen English teachers were hired via an agency each for one day. All teachers were native English speakers from the United States or Canada. One day is not enough to annotate the whole sentence set, therefore the TIMIT SX set was chosen for the ratings. We did not want to extend the evaluation over more than one day, as we would have expected a change in rating behavior on the next day and consistency of the ratings is an important issue. The SX set is phonetically compact, i.e. it is designed to cover as many biphone combinations as possible, therefore providing a good basis for examining pronunciation errors in phonetic contexts. Annotating the SX set took around 6 h. Each evaluator had to listen to 1,152 utterances (48 TIMIT sentences times 24 non-native speakers) in order to assign a utterance-level rating from 1 (best) to 5 (worst) in terms of pronunciation and fluency to each utterance and mark any words which are mispronounced. In total the speech data of 96 non-native speakers was evaluated, i.e. since there are 15 evaluators and each evaluator annotated the data of 24 non-native speakers, each speaker is assessed by three to four evaluators.

Each evaluator was given a comprehensive instruction sheet. The ratings were conducted on a web browser based interface. In the beginning, each evaluator had to listen to a uniform set of 22 calibration sentences. This set consists of 22 different sentences of 22 different non-native speakers. The evaluator had to assign a rating from 1 to 5 to each calibration sentence considering pronunciation and fluency. The selection criterion for the speakers was their recognition rate in the hotel reservation tasks TAS22001 and TAS32002 with a native acoustic model. For each of the five major first languages, one speaker with the highest, one

Table 5.4 Distribution of the expert pronunciation quality ratings at the word and the sentence level

Level	Sentence					Word
Label	1	2	3	4	5	Mispron.
Freq.	382 8.3%	1791 38.9%	1744 37.8%	623 13.5%	68 1.5%	3528 9.3%

speaker with the lowest and one speaker with a medium recognition rate were selected. Furthermore, one sentence of one speaker each of the remaining first languages (Bulgarian, Spanish, Portuguese, Hindi, Sinhalese, Hungarian and Korean) was included. Since the evaluators were asked to assign level 1 to the best utterances and level 5 to the worst utterances and to use each level at least once, they had an anchor for future ratings. The target of this calibration was to induce consistent rating behavior. The complete calibration set was accessible on one browser page, so that the evaluators could correct their earlier ratings after comparing to later utterances.

In the next step 48 sentences of 24 speakers were presented separately to the evaluator. The order of presentation was determined randomly and was different for each evaluator. The evaluator had first to listen to an utterance, mark mispronounced words and finally select a level of proficiency. For these utterance level ratings, the evaluators were instructed to ignore sentence intonation, for marking of words to consider phonetic pronunciation errors but to ignore wrong lexical stress. The evaluator was not allowed to go back to already processed utterances and change previously assigned labels. Table 5.4 shows the distribution of the marks 1 (very good English) to 5 (unintelligible pronunciation) and the number of words that were perceived as mispronounced (binary label).

To measure to what extend the evaluators agreed, and thus how reliable the ratings are, two types of measures can be employed:

- Pairwise correlation
- Open correlation

The equations for the two correlation measures are explained in Sect. 6.3.3.

The pairwise correlation on utterance level ranges between the rather low value of 0.28 and the acceptable value of 0.65, with an average of 0.49. Comparing to other work in literature [Franco 00], where an average correlation of 0.65 is reported, the correlation seems somewhat low. There are several possible reasons:

- The English teaching experience of the evaluators varies, with some teachers even having phonetic background knowledge, and others only on-the-job training as teachers.
- The interpretation of the term “mispronounced” and the marks 1 to 5 is subject to individual opinions. 22 utterances as evaluator training set might have been too short. For example, one evaluator reported he changed his behavior during

the experiment. On the other hand, having a longer calibration set would have stretched the ratings over more than one day, and the evaluators might have required another calibration on the next day.

- The evaluation strictness was very variant among the evaluators. For example, the pair with the lowest pairwise correlation consists of the evaluator who marked the most words as mispronounced and the one who marked the least.

The second measure is the so-called open correlation [Franco 00], which is the correlation between the mean rating of all but one evaluator and the remaining evaluator. The utterance-level inter-rater open correlation is between 0.45 and 0.70. Averaging the correlation values for all evaluators results a mean of 0.60.

The open correlation can also be calculated on speaker level. For each evaluator and speaker, the mean of the utterance-level ratings is computed. Calculating the inter-rater open correlation on these means leads to values between 0.88 and 0.98, with an average of 0.94.

This averaging out of individual evaluator effects shows that the ratings are sufficiently stable to be used as reference for research in automatic pronunciation scoring or computer aided language learning systems. In Appendix D we give detailed information about the rating procedure, in [Cincarek 04a] we further discuss inter- and intra-rater consistencies.

Table 5.5 shows a list of the words marked most often as mispronounced by at least one evaluator. The phonetic transcription is given in the IPA alphabet. Apparently, words with an *x* are hard to pronounce for any non-native speaker. On the other hand, the word *rare* is hardly a problem for German speakers of English, whereas most Chinese and Japanese speakers failed to pronounce it correctly. The ranking takes word frequency into account, looking only at the absolute number of error marks, the word "*the*" is the most frequently mispronounced word.

There is also a clear relation between length of a word and its chance to be marked as mispronounced. As Fig. 5.1 shows, the higher the number of phonemes in a word, the higher the relative marking frequency. This seems easy to understand, since an evaluator may mark a word if there is at least one mispronounced phoneme and the possibility for mispronunciation increases with the number of phonemes.

5.2.5 Comparison to Existing Databases

The ATR non-native database is a large step ahead for research on non-native speech. With almost 100 speakers and 15000 utterances, it is far larger than most non-native databases containing only small numbers of subjects and spoken sentences. Only large databases allow for a division in development and test set and still lead to statistically relevant results.

In order to examine the generality of data-driven methods it is important to have comparable data of several non-native accents. The ATR non-native database

Table 5.5 Words with a high mispronunciation frequency

Word	Pronunciation	All	German	French	Indonesia	China	Japan
Extra	[ɛkstɹə]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Exposure	[ɪkspoʃɚ]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Exam	[ɪgzæm]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Box	[bɑks]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Mirage	[məɹɑʃ]	**0.92**	0.71	**1.00**	**0.94**	**0.94**	**0.92**
Centrifuge	[sɛntɹɪfjʊdʒ]	**0.85**	**0.93**	**0.86**	**0.94**	**0.89**	0.72
Bugle	[bjʊgəl]	**0.85**	0.64	**1.00**	**1.00**	**0.89**	0.72
Frantically	[fræntɪkli]	**0.84**	0.79	**0.81**	**0.88**	0.67	**0.96**
Purchase	[pɝtʃəs]	0.76	0.64	**0.94**	**0.81**	0.67	0.76
Rare	[ɹeɪ]	0.75	0.36	0.69	0.75	**0.89**	**0.88**
Contagious	[kəntedʒəs]	0.74	0.57	0.69	**0.81**	**0.83**	0.72
Formula	[fɔɹmjələ]	0.73	0.79	**0.88**	**0.81**	0.67	0.56
Ambulance	[æmbjələns]	0.73	0.64	**0.81**	0.75	0.78	0.72
Development	[dɪvɛləpmənt]	0.70	0.36	**0.88**	**0.94**	0.67	0.64
Pizzerias	[pitsɚiəz]	0.69	0.36	0.56	**0.88**	0.78	0.76
Guard	[gɑɹd]	0.69	0.43	0.75	0.75	**0.83**	0.68
Colored	[kʌlɚd]	0.69	0.50	**0.81**	0.56	0.67	**0.84**
Chablis	[ʃəbli]	0.69	0.36	0.44	**0.88**	**0.83**	**0.80**
Thursdays	[θɝzdez]	0.68	0.50	0.69	0.75	0.67	0.76
Mergers	[mɝdʒɚz]	0.67	0.29	0.69	**0.81**	0.72	0.72

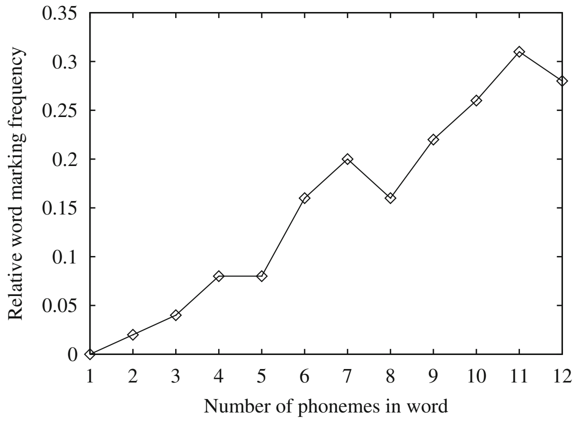


Fig. 5.1 Relationship between the number of phonemes in a word and its relative marking frequency

includes five large accent groups of non-native English. Other major databases like [Byrne 98, Minematsu 02] focus on a single accent. Our database has precise information about speaker properties, especially native language, which is necessary to train accent-dependent models. This information is missing in other databases [Benarousse 99, van den Heuvel 06].

Skill and mispronunciation annotation is a prerequisite to conduct research in automatic pronunciation scoring for CALL systems. Our database provides both a sentence-level rating as well as mispronounced word tags. Other corpora with such labels focus only on a single accent [Minematsu 04] or on learners of Japanese [Minematsu 02] which makes studies difficult for non-Japanese researchers.

Non-nativeness is already a challenging problem for speech recognizers, if overlapped by strong noise or extreme spontaneousness, speech recognition accuracy becomes so low that random effects can influence blur performance measures. Some databases that are large but contain such additional challenges [Lamel 94, Teixeira 97, Pigeon 07] find little use in the research community. We took special care in order to avoid such overlapping in the design of our recording environment.

Finally a database is only helpful if it is possible to acquire at least a research license. The otherwise very interesting Cross-Towns corpus [Schaden 02] unfortunately remains closed property of the collecting institute, with no release planned for the near future. The ATR non-native database can be licensed at ATR, a well-known source of speech databases in Japan.

Chapter 6
Handling Non-native Speech

Other than the main contribution of this work, the HMMs as statistical pronunciation models we describe in Chap. 7, we have conducted several experiments to handle non-native speech. These include rule-based phoneme lattice processing and multilingual weighted codebooks of semi-continuous HMM based recognizers. Section 6.3 gives a brief outlook into automatic scoring of non-native speakers pronunciation and mispronounced word detection, comparing a wide range of features on sentence and word level. Automatic scoring algorithms are the basis for computer assisted language learning (CALL) systems. Finally, we look into the prosodic analysis of non-verbal utterances in Sect. 6.4 as they can contribute to a deeper understanding of non-phonemic effects. In the following, we will describe those experiments.

6.1 Rule-Based Lattice Processing

An interesting rule-based approach to compensate non-native pronunciation errors is processing multilingual phoneme lattices [Gruhn 02, Binder 02]. The goal of this technique is to correct variations in pronunciation at the phoneme level. As we analyze specifically Japanese speakers of English, we take into account the possibility that Japanese speakers use phonemes more similar to Japanese than to English by applying a bilingual acoustic model. To keep as much information of the original speech as possible, all phonemes of the English and the Japanese language are included into a bilingual HMM acoustic model. Consonants that share the same IPA-symbol are represented only once, to reduce the confusion. According to rules all Japanese phonemes are matched on English phonemes and additional variations for English phonemes are added, too. Comparable Methods at the word level [Tomokiyo 00] have already achieved good results, but if a word is misrecognized because of differences in pronunciation, it is lost for further

R. E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*,
Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_6,

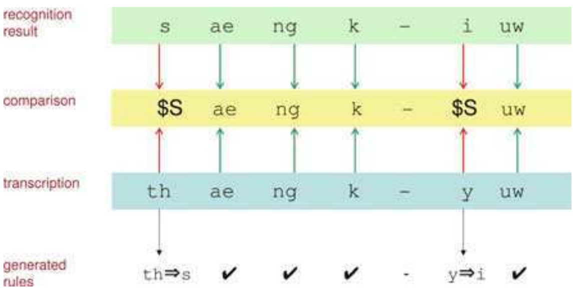


Fig. 6.1 Overview on rule derivation from n-best result, transcription DP-alignment for "thank you". $\$S$ marks substitutions and $\$I$ insertions. Japanese phonems are written in lower case and English phonemes in capital letters

processing. By adding the pronunciation variations to the phoneme-lattice, this new approach allows a more robust recognition, without putting many variations of the same word into the dictionary.

Training data for the native English acoustic model (AM) was the training set of the WSJ database and conducted as described in Sect. 7.5.1, except that the software ATR-SPREC [Singer 99b] was used for acoustic model training and recognition. The Japanese AM training was conducted on the ATR's Japanese TRA-Database [Nakamura 96], a read speech database in a hotel reservation scenario. The separately trained English and Japanese models were merged into one mixed HMM set.

The non-native database for both rule generation and test was a part of the ATR non-native English database described in Chap. 5 containing only the first eleven Japanese speakers of English. In this subset, we have 1342 utterances for training and 244 for test. The division into test and training sets is identical to the choice given there, the test set consists of two hotel reservation dialogs. Evaluations were based on the smaller dictionary with 6,700 pronunciations for 2,700 words.

6.1.1 Rule Generation

An unrestricted 1-best phoneme-recognition was performed on the non-native training data and the result matched to the transcription by dynamic programming. As shown in Fig. 6.1 the results are the recognized phonemes and a sequence, containing the phoneme when recognized correctly, $\$I$ for insertion and $\$S$ for substitutions. Deletions would be marked with $\$D$. The substitutions are the basis information for the rule generation.

The variations in pronunciation extracted have to be transformed from an unsorted order into rules containing information about the frequency of appearance. The number $N_{ph_{rec} \to ph_{corr}}$ of each variation $ph_{rec} \to ph_{corr}$ is counted and divided by the frequency $N_{ph_{rec}}$ of the phoneme ph_{rec}.

$$P_{sub} = P(ph_{i,corr}|ph_{j,rec}) = \frac{N_{ph_{i,rec} \rightarrow ph_{j,corr}}}{N_{ph_{i,rec}}}. \tag{6.1}$$

For this calculation, only correct recognized phonemes and substitutions, but not the insertions and deletions were taken in account. All rules were added to the list, without any limitation to probability [Amdal 00a]. A selection of rules more frequent than a given probability threshold T_{sub} takes place during application though. It is important to notice that, unlike in [Witt 99b], there is no mapping of a recognized phoneme to one target phoneme. The original phoneme is substituted by likely variations, which are only from the English phoneme set and also can contain the original phoneme.

6.1.2 Lattice Processing

Figure 6.2 explains the lattice processing algorithm. It is applied on the phoneme lattice created by the decoder (Fig. 6.3). All Japanese phonemes are substituted with English variations. If rules for an English phoneme are existent, they are also going to be applied. In this case, the phoneme is substituted by its variations, which can contain the phoneme itself, too. The log-likelihood for these new possible phonemes is calculated using the log-likelihood $L(ph_{orig})$ of the original phoneme and the previous calculated probability $P(ph_{rec}|ph_{corr})$ for the variation.

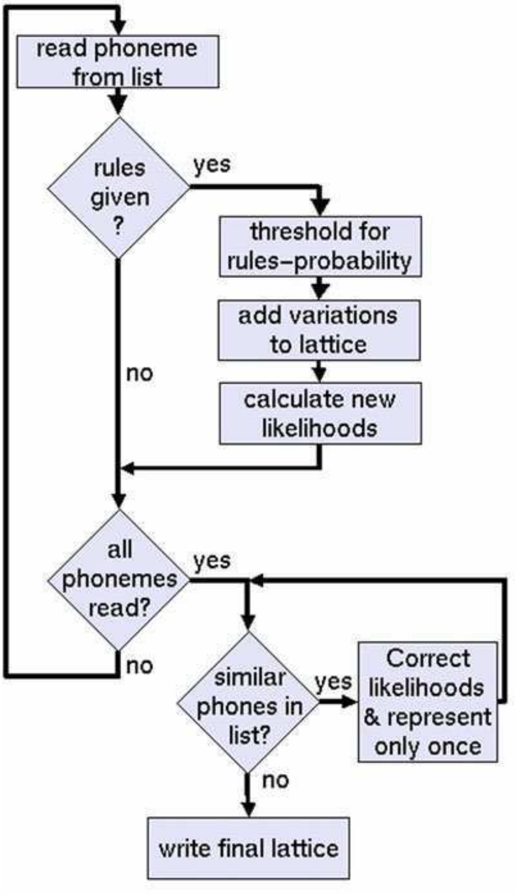
$$L(ph_{new}) = L(ph_{orig}) + log(P(ph_{rec}|ph_{corr})) \tag{6.2}$$

After the new variations are included, a phoneme might appear more than once between the same nodes. In this case, it is dynamically reduced to one arc and the likelihoods of this arc is adjusted. The resulting lattice has more arcs than the original but the same number of nodes and contains only phonemes of the English phoneme-set.

We evaluated the performance depending on the number of rules. Each rule has a probability P_{sub} measuring the occurrence frequency during training, we select only those rules with P_{sub} larger than a threshold T_{sub}, experimenting with values from 3 to 11%. The total number of rules in the rule-set was 1298. Including rules with lowest occurrence worsened the result as random phoneme confusions by recognition errors during rule generations start to have a large impact. In Table 6.1 we show two measures to evaluate the performance, phoneme net accuracy (PNA) and phoneme best correctness (PBC). PNA is the accuracy for the best path through the phoneme lattice, ignoring the score. PBC is the correctness rate of the 1st best result with the highest acoustic score.

As it can be seen in Table 6.1, the phoneme net-accuracy has a maximum for a threshold of 7% (80 rules). If more rules are applied, the recognition suffers from

Fig. 6.2 The lattice processing algorithm


read phoneme from list
rules given ?
yes
threshold for rules–probability
add variations to lattice
calculate new likelihoods
no
all phonemes read?
yes
no
similar phones in list?
yes
Correct likelihoods & represent only once
no
write final lattice

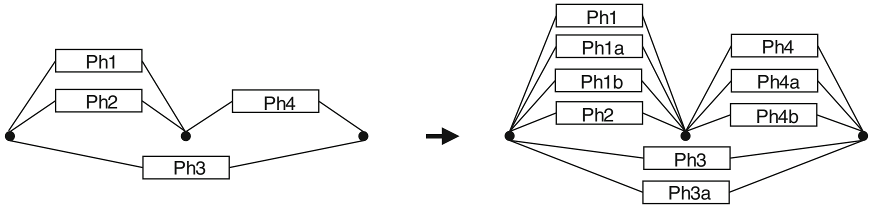


Fig. 6.3 An example for a lattice before (*left*) and after processing (*right*)

confusion, for fewer rules, the recognition result converges to the recognition rate of the unprocessed lattice. The overall phoneme best correct rate for the processed lattice is lower than for the unprocessed lattice. As PNA results prove, the correct phonemes do exist in the lattice. This may be indicating that the calculation of the score during variation application is not yet optimal.

Table 6.1 Phoneme lattice processing with variable sized rule sets, depending on occurrence frequency

T_{sub}	Unpr	0.03	0.07	0.08	0.10	0.11
# Rules	—	236	80	65	50	44
PNA	57.3%	64.7%	65.9%	64.4%	62.0%	60.9%
PBC	34.5%	30.1%	31.5%	31.5%	32.5%	32.3%

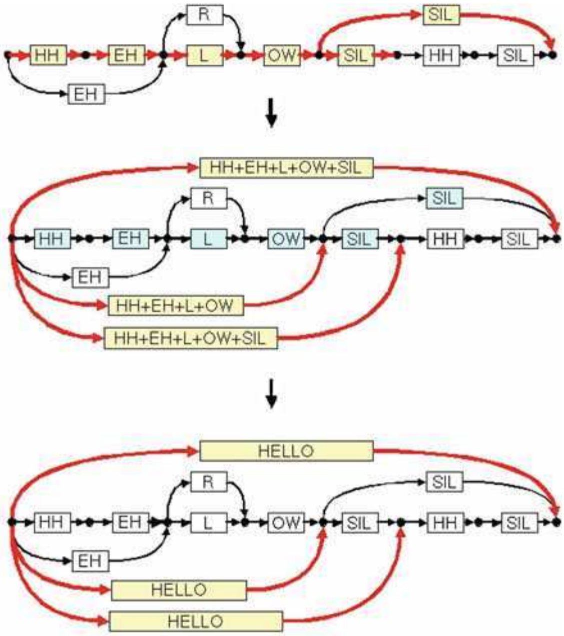


Fig. 6.4 Example for a conversion from a phoneme to a word lattice

6.1.3 Extension to Word Level

In order to get an estimate how well the method performs not only for phoneme recognition, we extended the method by transforming the phoneme to a word level as shown in Fig. 6.4. As a phoneme lattice tend to have in the order of 10^3 nodes and the dimension of arcs is 10^6 this problem does not seem to be solvable in reasonable time. For the calculation of the word correct rate, only the number of correct words as defined in the transcription is necessary. Thus, the dictionary-lookup was restricted to those entries. With this evaluation method, we achieved an improvement from 73.0% to 82.3% word correctness. But these numbers must be seen with some caution, as e.g. in the example shown in Fig. 6.4 a misrecognition from the word hello to low (both including the phoneme sequence /L+OW/) is not possible.

Table 6.2 Recognition results (word accuracy) comparing a baseline pronunciation dictionary and dictionaries including added pronunciation variations

T_{sub} Dict size	Unpr 6664	0.08 41151	0.10 23506	0.15 14218	0.17 12142	0.20 9994
Total	23.24	25.00	25.95	25.90	26.49	25.81
Good	48.12	42.08	46.59	48.01	49.64	48.19
Medium	29.06	30.77	31.62	32.30	33.16	31.62
Poor	12.48	20.51	19.49	17.26	17.43	16.75
Bad	2.91	5.24	5.23	5.97	5.52	5.47

6.1.4 Applying Rules on the Pronunciation Dictionary

Alternatively to lattice processing, the phoneme confusion rules can also be applied to add pronunciation variants to a standard dictionary. During lattice processing, we could take the probability of a rule directly into account by adjusting the arc likelihood and therefore theoretically apply all rules given. A standard pronunciation dictionary does not contain such probabilities, and therefore we cannot apply all possible rules to the dictionary. Especially as frequently not only one, but several rules can be applied to a word and the number of possible permutations explodes for long words. Similarly to the lattice processing, we consider only those rules, whose probability P_{sub} is larger than a defined threshold T_{sub}. The results for various choices of T_{sub} and therefore number of entries in the dictionary are shown in Table 6.2. Non-nativeness is a term covering a wide range of skills. To examine performance for good and bad speakers, the test set was divided into four groups of equal size with the baseline performance of each speaker as clustering criterion.

While the experiments in this section show that rule-based phoneme lattice parsing can improve recognition accuracy, the gain is only small or existing but impractical to apply in a real world system. A general problem about rule-based approaches is that errors other than substitutions are difficult to handle - in fact, in this approach, they are not considered at all. This limits the performance of such approaches.

6.2 Multilingual Weighted Codebooks

For the special case of semi-continuous HMMs, multilingual weighted codebooks [Raab 08a] have proven to be an effective method to increase performance for non-native speech recognition [Raab 08b, Raab 08c].

Semi-continuous systems are based on a vector quantization (VQ). As the codebooks for VQ are calculated on the training data of the acoustic model, they only fully represent the properties of the one language they were made for and perform much worse for speech from other languages. Semi-continuous HMMs

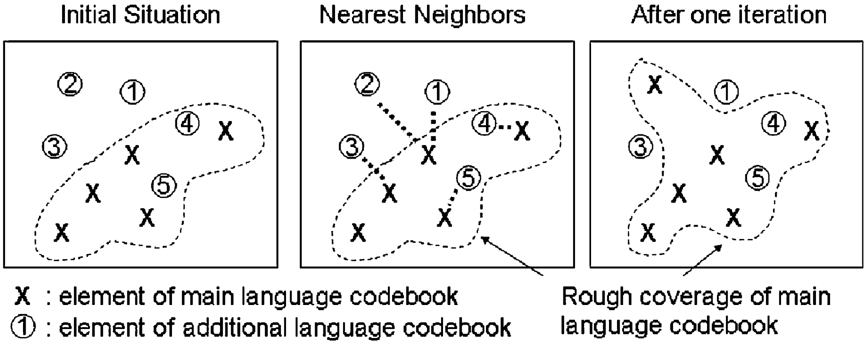

Initial Situation
Nearest Neighbors
After one iteration
X : element of main language codebook
① : element of additional language codebook
Rough coverage of main language codebook

Fig. 6.5 Basic idea of multilingual weighted codebooks

are typically applied in limited-resource tasks such as car navigation devices and have to optimally support a main language for commands and navigation, but also recognize some additional languages e.g. for input of international destinations. Resource reasons prevent simply multiplying the codebook size, and the focus on the main language forbids creating a fully multilingual codebook. The multilingual acoustic model, on the other hand, can be a simple concatenation of several monolingual models as not all languages are required at the same time.

To achieve the first priority aim of keeping the main language accuracy, the multilingual weighted codebooks (MWCs) created always contain all Gaussians from the main language codebook. The Gaussians that originate from the main language always remain unchanged. To improve the performance on the additional languages, the MWCs will also contain some Gaussians which originate from codebooks from the additional languages, with one additional Gaussian being added per iteration. After the multilingual weighted codebook is generated, the acoustic model has to be retrained with the new codebook.

The proposed MWC is therefore the main language codebook plus some additional Gaussians. Figure 6.5 depicts an example for the extension of a codebook to cover an additional language. From left to right one iteration of the generation of MWCs is represented in a simplified two dimensional vector space.

The picture to the left shows the initial situation. The X's are mean vectors from the main language codebook, and the area that is roughly covered by them is indicated by the dotted line. Additionally, the numbered O's are mean vectors from the second language codebook. Supposing that both X's and O's are optimal for the language they were created for, it is clear that the second language contains sound patterns that are not typical for the first language (O's 1, 2 and 3).

The picture in the center shows the distance calculation. For each of the second language codebook vectors, the nearest neighbor among the main language Gaussians is determined. These nearest neighbor connections are indicated by the dotted lines.

The right picture presents the outcome of one iteration. From each of the nearest neighbor connections, the largest one was chosen as this is obviously the mean vector which causes the largest vector quantization error. In the pictures, this is O number 2. Thus the Gaussians from O number 2 was added to the main language codebook.

Experiments were conducted on the ISLE corpus [Menzel 00] of non-native English. The results show it is possible to get the same performance with only 10% additional codebook vectors as with concatenating both codebooks. The overall improvement (either with a double codebook or 10% increased codebook size) has been a relative 10% compared to a monolingual codebook.

Further optimization of the approach is possible in an on-demand system [Raab 11]: After analyzing the required language combinations for a given application in a specific context, a multilingual acoustic model can be created on-the-fly, based on multilingual codebooks. While such generated acoustic models perform inferior to properly trained models, they still have the advantage of accomodating even exotic accents.

6.3 Automatic Scoring of Non-native Pronunciation Skills

To learn a second language without a human teacher, computer assisted language learning (CALL) systems are very helpful. While grammar and vocabulary can be learned in self-study, for pronunciation some kind of correctness judgement is needed. From a pedagogical point of view, a system for computer assisted pronunciation training should provide the student an overall assessment of pronunciation quality to verify correctness, pinpoint certain rather than highlight all mistakes, and possibly suggest a remedy [Neri 04]. In order to provide feedback without human assistance, methods for automatically scoring the pronunciation quality at different levels of granularity are required.

In this research [Cincarek 09, Cincarek04b], we present a pronunciation scoring method applicable independently of the non-native's first language. We provide a method to derive feedback about mispronunciations at the phoneme level from word level scoring.

The method has been evaluated on the database explained in Chap. 5. As only for the TIMIT phonetically compact sentences human skill ratings are available, only those sentences were the basis for pronunciation scoring. The experimental setup is as described in Sect. 7.5: A monophone acoustic model trained on the WSJ database is used for phoneme and word recognition, combined with a hotel reservation task language model.

In [Cincarek 09], we propose the following features for automatic pronunciation scoring:

6.3.1 Sentence Level Pronunciation Scoring

In Table 6.3 variables and symbols used in feature definitions are summarized. Feature extraction is carried out separately for each sentence **S**. A sentence can be

Table 6.3 Definition of variables and symbols for sentence level pronunciation features

Entity	Symbol	Definition
Sentence	$\mathbf{S}$	Word sequence $(W_1, \ldots, W_M)$
		Phoneme sequence $(p_1, \ldots, p_N)$
		Phoneme segments $(\mathbf{X}_1, \ldots, \mathbf{X}_N)$
Segment	$\mathbf{X}$	Frame sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$
Frame	$\mathbf{x}$	Acoustic features $(x_1, \ldots, x_d)$
Duration	T	Phoneme segment duration ($\mathbf{X}$)
	D	Word segment duration (W)
	T_S	Total sentence duration
# Phonemes	N	No. of phoneme segments in $\mathbf{S}$
# Words	M	No. of word segments in $\mathbf{S}$
Speaking rate	$R^{(ph)}$	# Phonemes (N) / Time (T_S)
	$R^{(wd)}$	# Words (M) / Time (T_S)

said to consist of either N phoneme segments or M word segments, which are also made up of phoneme segments. It is assumed, that there are no intra-word pauses, but only inter-word pauses. The duration of a phoneme segment $\mathbf{X}_i$ with label p_i is denoted as T_i. Word durations are denoted as D_j. The total duration T_S of a sentence is defined as the duration of all phonemes plus inter-word pauses in the sentence. Leading and trailing silence segments are ignored. The rate of speech (ROS) is a measure of the speaking rate. It can be defined as the number of phonemes, syllables or words per time.

The rate of speech can be used as pronunciation feature. However experiments revealed that there is a higher correlation for the reciprocal phoneme-based rate of speech, i.e. the mean phoneme duration:

$$\text{(MeanPhDur)}\, \mathcal{R} = \frac{1}{R^{(ph)}} \tag{6.3}$$

Another feature is the duration score [Neumeyer 00] to measure deviations from the duration characteristics typical for native speech. The duration log-likelihood of the phoneme models in the sentence is summed up as follows:

$$\text{(DurScore)}\, \mathcal{D} = \frac{1}{N} \sum_{i=1}^{N} \log P_{dur}^{(ph)}(T_i * R^{(ph)} | p_i) \tag{6.4}$$

A phoneme duration probability density function (pdf) can be estimated from transcribed native speech data. Instead of approximating the pdf with a histogram, the log-normal density function

$$P_{dur}^{(ph)}(t|p) = \frac{1}{t\sqrt{2\pi\sigma_p^2}} \exp\left[-\frac{(\log t - \nu_p)^2}{2\sigma_p^2}\right] \tag{6.5}$$

is employed, since phoneme durations are distributed log-normal. The parameters v_p and σ_p are obtained by maximum-likelihood estimation based on ROS-normalized duration samples for each phoneme. This normalization is necessary in order to account for variations of the speaking rate.

The acoustic model likelihood $L(\mathbf{X}) = \log P(\mathbf{X}|\lambda_p)$ can be considered as a measure of acoustic similarity between the target speech and the context-independent acoustic model λ_p for phoneme p. Here, the original definition of the likelihood-based pronunciation feature [Neumeyer 00] is modified by additionally normalizing with the rate of speech, since the correlation to human ratings increased further.

$$\text{(SentLh1)}\, \mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \frac{L(\mathbf{X}_i)}{T_i * R^{(ph)}} \tag{6.6}$$

To calculate feature $\mathcal{L}$, each segment's likelihood is divided by its actual duration. Alternatively, normalization is possible by dividing with the expected (phoneme or word) duration. This is realized for the following new pronunciation feature:

$$\text{(SentLh2)}\, \mathcal{E} = \frac{1}{M} \sum_{j=1}^{M} \frac{L(W_j)}{D_j^{(e)} * R^{(wd)}} \tag{6.7}$$

$L(W_j)$ denotes the sum of phoneme model log-likelihoods of word W_j. An estimate for the expected word duration $D_j^{(e)}$ is the sum of the mean duration of the phonemes of word W_j.

Besides the phoneme likelihood, the phoneme posterior probability $P(p_i|\mathbf{X})$ is a promising pronunciation feature. In [Neumeyer 00] it was shown to be the feature with highest correlation to human ratings. Its calculation was simplified to the likelihood ratio

$$L_r(X_i|p_i) = \sum_{t=1}^{T_i} \log \frac{P(\mathbf{x}_t|p_i)}{P(\mathbf{x}_t|q_t^*)} \tag{6.8}$$

where q_t^* is the name of the model with highest likelihood given frame $\mathbf{x}_t$. In practice, q_t was obtained by unconstrained phoneme recognition. Thus a likelihood ratio score was obtained for each phoneme segment. These scores are normalized by the actual segment duration, summed up and finally divided by the number of segments N. Here, the feature is modified to

$$\text{(LhRatio)}\, \mathcal{K} = \frac{\sum_{i=1}^{N} L_r(\mathbf{X}_i)}{\sum_{i=1}^{N} T_i^{(e)} * R^{(ph)}} \tag{6.9}$$

i.e. normalizing the segments posterior scores by the product of the speaking rate and the expected segment duration $T(e)$, since the correlation to human ratings increased further.

An indicator of how well an utterance can be recognized is the phoneme or word accuracy. The former is a better measure, since it is based on a larger number of tokens. The accuracy can be calculated as the normalized minimum-edit-distance:

$$\text{(PhAcc)}\,\mathcal{A} = \frac{\mathit{MinEditDist}(\mathbf{q}, \mathbf{p})}{\max\{|\mathbf{q}|, |\mathbf{p}|\}} \tag{6.10}$$

The distances of insertions, deletions and substitutions are uniformly set to one. $|\mathbf{p}|$ means the number of phonemes in the phoneme reference vector $\mathbf{p}$. $\mathbf{q}$ denotes the phoneme recognition hypothesis. $\mathcal{A}$ is zero if reference and hypothesis are identical and greater than zero, if there are recognition errors.

Being unsure about a word's pronunciation may introduce inter-word pauses. Consequently, it is worth considering the total duration (PauseDur) $\mathcal{P}$ of inter-word pauses [Teixeira 00] within a sentence as a feature.

As a further new pronunciation feature the probability of the recognized phoneme sequence $\mathbf{q}$ given an n-gram language model (LM) is employed. The LM should be trained on canonic phoneme transcriptions of valid sentences of the target language, because a foreign language student should acquire standard pronunciation.

$$(\mathit{PhSeqLh})\,\mathcal{M} = \frac{1}{R^{(ph)}} \log P(\mathbf{q}|\text{LM}) \tag{6.11}$$

Each pronunciation feature is intended to measure certain aspects of pronunciation. $\mathcal{R}, \mathcal{D}$ and $\mathcal{P}$ are measures for temporal characteristics like the fluency of a speaker. $\mathcal{L}$ and $\mathcal{K}$ are intended to measure the segmental quality. $\mathcal{M}$ and $\mathcal{A}$ can be considered as indicators for both kinds of characteristics.

There are two approaches for sentence scoring that we examined. The Gaussian classifier with maximum likelihood decision rule itself can provide a discrete scoring result (hard scoring). A continuous scoring result (soft scoring) can be obtained by calculating the expected score value from the likelihood of the class models and the class prior probabilities. The latter are considered to be distributed uniformly. Another approach for soft scoring is to use a linear combination of the pronunciation features. The weighting coefficients for each feature can be estimated using linear regression.

6.3.2 Word Level Pronunciation Scoring

Any feature defined for the sentence level can be applied to the word level in principle, since sentences consisting of only one word are valid. However,

Table 6.4 Definition of variables and symbols for word level pronunciation features

Entity	Symbol	Definition
Word	$\mathbf{W}$	Word labels $(W_1, \ldots, W_M)$
Sequence	$\mathbf{O}$	Acoustic observ. $(O_1, \ldots, O_M)$
Word	W	Phoneme labels $(p_1, \ldots, p_n)$
	O	Acoustic segments $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$
Phoneme	$\mathbf{X}$	Frame sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$
segment		Reference labels $(p_1, \ldots, p_T)$
		Hypothesis labels $(q_1^*, \ldots, q_T^*)$
# Phonemes	n	No. of phonemes in word W

preliminary investigations revealed that features with high quality for the sentence level are not necessarily good for the word level. Table 6.4 briefly explains variables and symbols employed for feature definitions.

Instead of using a duration-normalized likelihood or the likelihood ratio, the plain sum of phoneme log-likelihoods $\mathcal{W}_1$ had a higher discriminative ability. Normalization of this feature is possible by dividing with the number of phonemes n in each word.

$$\text{(WLh1)}\, \mathcal{W}_1 = \sum_{i=1}^{n} L(\mathbf{X}_i) \quad \text{(WLh2)}\, \mathcal{W}_2 = \frac{1}{n}\mathcal{W}_1 \tag{6.12}$$

The sentence duration score $\mathcal{D}$ is a good word level feature without modifications:

$$\text{(DurS1)}\, \mathcal{W}_3 = \sum_{i=1}^{n} S_{dur}^{(ph)}(T_i * R^{(ph)} | p_i) \tag{6.13}$$

The following normalizations of $\mathcal{W}_3$ were advantageous in some cases:

$$\text{(DurS2)}\, \mathcal{W}_4 = \frac{1}{n}\mathcal{W}_3 \quad \text{(DurS3)}\, \mathcal{W}_5 = \mathcal{W}_3\mathcal{R} \tag{6.14}$$

Confidence measures showed to have the highest discrimination ability. The feature $\mathcal{C}_1$ is a high-level confidence measure derived with the phoneme correlation technique from [Cox 02]. It is based on the phoneme confusion matrices for correctly pronounced and mispronounced words. The confusion probabilities are calculated at the frame level. As for the calculation of the likelihood ratio in Eq. 6.8 q_t^* denotes the phoneme label of the speech frame derived from unconstrained phoneme recognition. The label p_t is obtained from the forced-alignment.

$$\text{(PhCfRatio)}\, \mathcal{C}_1 = \frac{1}{D}\sum_{t=1}^{D} \log \frac{P(q_t^* | p_t, \text{wrong})}{P(q_t^* | p_t, \text{correct})} \tag{6.15}$$

Another confidence measure is the word posterior probability (WPP) [Wessel 01]. It measures the degree to which a word recognition hypothesis can be trusted.

It may be assumed, that the value of the WPP also reflects the pronunciation quality of a word. The word level pronunciation feature $\mathcal{C}_2$ is based on the sentence likelihood. It was calculated via N-best lists in order to be independent from the architecture and implementation of a speech recognizer.

$$(WPP)\,\mathcal{C}_2 = \frac{\sum_{\mathbf{V}} P(\mathbf{O}|\mathbf{V}) f(W_j|V_i)}{\sum_{\mathbf{V}} P(\mathbf{O}|\mathbf{V})} \tag{6.16}$$

The summation is carried out over the word sequences $\mathbf{V} = (V_1, V_2, \ldots, V_i, \ldots)$ of each hypothesis from the N-best list. The function $f(W_j|V_i)$ returns 1, if the overlapping condition for the reference word W_j and a word V_i in the hypothesis is met. Otherwise its value is 0. The language model probability $P(\mathbf{V})$ is not employed for the calculation of the WPP, since the feature should only be based on acoustic evidence.

Mispronounced words could be detected using a continuous word score as in sentence scoring and a threshold to decide on mispronunciations. Since the purpose is in the end to discriminate correctly pronounced words (correct) from mispronounced words (wrong), the issue is a two-class classification problem. For the discrimination of the two classes of correctly pronounced and mispronounced words the Gaussian classifier is employed. Other methods for classification, decision trees (CART) and Gaussian Mixture Models (GMMs) after reduction of the feature space dimension with principal component analysis (PCA) could not outperform the Gaussian classifier.

6.3.3 Scoring Experiments

To evaluate the scoring system, we assigned skill scores to the phonetically compact sentences from the ATR database. To measure the performance of the scoring system, the correlation coefficient *C(X,Y)* is employed. It is defined as

$$C(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2 \sum_{i=1}^{n}(y_i - \mu_Y)^2}} \tag{6.17}$$

where the values x_i, y_i of the random variables *X, Y* are corresponding pronunciation annotations with index *i* assigned by the human labelers and the automatic scoring system, respectively.

The performance for sentence scoring with the Gaussian classifier and linear transformation based on single features is shown in Table 6.5. The table shows the correlation to the human ratings. The best four single features are the likelihood ratio $\mathcal{K}$ followed by the phoneme accuracy $\mathcal{A}$, the duration score $\mathcal{D}$ and the likelihood score $\mathcal{E}$ normalized by the expected word duration.

Table 6.5 Sentence scoring based on single pronunciation features sorted by correlation to human ratings

Feature	Correlation
$\mathcal{K}$ (LhRatio)	0.50
$\mathcal{A}$ (PhAcc)	0.44
$\mathcal{D}$ (DurScore)	0.42
$\mathcal{E}$ (SentLh2)	0.40
$\mathcal{L}$ (SentLh1)	0.38
$\mathcal{M}$ (PhSeqLh)	0.38
$\mathcal{R}$ (MeanPhDur)	0.35
$\mathcal{P}$ (PauseDur)	0.31

It is possible to combine those features with linear combination. The results are shown in Table 6.6. The lower three sets include features which can be calculated given only the forced-alignment. There is a remarkable increase in performance if the features based on the recognition result are also employed (upper three sets).

Table 6.7 shows the positive effect of score adjustment. Scoring results '1' or '5' almost never occur. In total, the sentence level scoring with the feature set $\mathcal{K}, \mathcal{A}, \mathcal{M}, \mathcal{D}$ leads to a class recognition rate of 86.6%.

For sentence level scoring, the goal of an automatic system is to be as close as possible to the ratings of human experts. For the binary problem of word pronunciation correctness, special attention must be payed on the type of error. In a real system, not pointing out a mispronounced word is much less a problem than scolding a student for a word he has pronounced correctly. Therefore, for word pronunciation analysis, we calculate the weighted recognition rate (WRR): The error of misclassifying a correctly pronounced word as mispronounced is weighted three times higher than falsely accepting a mispronounced word as correct. Table 6.8 shows the WRR for discriminating correctly pronounced words from mispronounced words based on single pronunciation features. The best two features are the phoneme confusion ratio $\mathcal{C}_1$ and the word likelihood $\mathcal{W}_1$.

As for sentence scoring, the *n*-best feature sets are determined heuristically with the floating search algorithm. Table 6.9 shows the combination leading to the best recognition rate. There was no significant increase in performance if employing five or more features.

Measures that are based on phoneme recognition accuracy or related scores have shown a high relevance for pronunciation scoring. Figure 6.6 illustrates the correlation between phoneme error rate of the speech recognizer and human rating. The rates were calculated over the TIMIT SX sentence set. As this type of text is comparatively difficult to pronounce, the phoneme error rates can exceed 100% for the lower-skill speakers. It is clearly visible that speakers who received a high rating by human evaluators also scored lower phoneme error rates by the speech recognizer.

The reliability of word level mispronunciation detection can be assessed when comparing the confusion matrix of human evaluators with the classifier's confusion matrix. To obtain the former, the majority voting of all but one evaluator was taken as reference and tested against the decision of the remaining evaluator.

Table 6.6 Result for sentence scoring by linear combination of multiple pronunciation features

Feature	Correlation
$\mathcal{K}, \mathcal{A}, \mathcal{M}, \mathcal{D}$	0.59
$\mathcal{K}, \mathcal{M}$	0.58
$\mathcal{K}, \mathcal{A}$	0.55
$\mathcal{D}, \mathcal{E}$	0.52
$\mathcal{D}, \mathcal{L}, \mathcal{R}$	0.48
$\mathcal{D}, \mathcal{L}$	0.47

Table 6.7 Confusion matrix obtained after rounding scores with feature combination $\mathcal{K}, \mathcal{A}, \mathcal{M}, \mathcal{D}$

	1	2	3	4	5
1	**52**	35	12	1	0
2	25	**39**	24	10	2
3	11	27	**34**	20	9
4	1	8	21	**31**	39
5	0	0	4	33	**63**

Table 6.8 Weighted recognition rate (WRR) for word classification with a Gaussian classifier based on single pronunciation features for the word level

Feature	WRR
$\mathcal{C}_1$ (PhCfRatio)	66.6
$\mathcal{C}_2$ (WPP)	66.0
$\mathcal{W}_1$ (WLh1)	65.8
$\mathcal{A}$ (PhAcc)	64.7
$\mathcal{M}$ (PhSeqLh)	64.1
$\mathcal{W}_3$ (DurS1)	64.0
$\mathcal{W}_5$ (DurS3)	61.2
$\mathcal{W}_4$ (DurS2)	58.0
$\mathcal{W}_2$ (WLh2)	54.5

Table 6.9 Weighted recognition rate with the Gaussian classifier based on multiple pronunciation features

Features	WRR
$\mathcal{W}_1, \mathcal{C}_2$ (WLh1, WPP)	70.7
$\mathcal{W}_1, \mathcal{C}_2, \mathcal{W}_4$ (+DurS2)	71.6
$\mathcal{W}_1, \mathcal{C}_2, \mathcal{W}_4$, $\mathcal{C}_1$ (+PhCfRatio)	72.2
$\mathcal{W}_1, \mathcal{C}_2, \mathcal{W}_4, \mathcal{C}_1, \mathcal{W}_2$ (+WLh2)	72.1

The average performance of four reference and test combinations is shown in the Table 6.10. There is a disagreement about 8% of the correct words and 42% of the wrong words. From the left table it is clear, that the detection of mispronounced works equally well by automatic classification. However, at the same time the classification error for correctly pronounced words is about 10% higher than for the human evaluators.

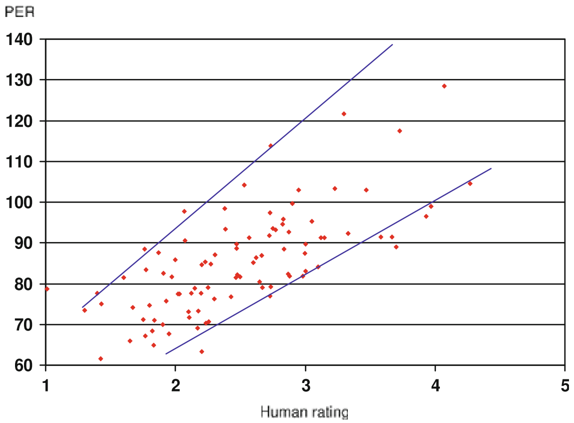


Fig. 6.6 Correlation between phoneme error rate and human rating

Table 6.10 Comparison of the confusion matrices of the human evaluators and the automatic scoring

	Correct	Wrong
Machine		
Correct	82.9	17.1
Wrong	42.1	57.9
Humans		
Correct	91.9	8.1
Wrong	42.2	57.8

In conclusion the proposed method to score the pronunciation skill of non-native speakers was almost as reliable as human expert ratings. For an estimation of pronunciation skill on sentence level, phoneme sequence probability and durations score was the most successful feature combination. The word posterior probability and phoneme confusion probability ratio of correctly pronounced and mispronounced words were identified as new word level features that contribute greatly to recognition rates, leading to a total weighted recognition rate of 72%. Perfect detection remains difficult though, and even human evaluators show some level of disagreement.

6.4 Classification of Non-Verbal Utterances

Listening to non-native speech, it is common to perceive an increased occurence of hesitations, fillwords, non-verbal expressions, duration variation and other prosodic effects. The most likely reason is insecurity: Speakers have to think about how to pronounce words, and at the same time prefer shorter common words which they have encountered with relative frequency and know how to

Table 6.11 Label number and according speech act

Number	Speech act
1	Listen
2	Understand
3	Interest
4	Affirm
5	Affirm overall
6	Call back
7	Disagree
8	Emotion

employ. Therefore automatic classification of common fillwords and non-verbal expressions can be seen as an important part of non-native speech recognition research.

Fillwords, or filled pauses, are common in most languages. They are transcribed with words like “well” or “ah” in English, “äh” in German or “un” in Japanese. Depending on their prosodic properties they can carry a wide range of meanings. In a dialog they take the role of providing a conversational backchannel giving feedback from affirmation to disagreement. This backchannel information is very relevant as one of its functions is to express understanding of the previous utterance, which is crucial for conversation with a non-native speaker. Fillwords can also express emotions.

In this section, we analyze how to classify non-verbal utterances by prosodic means in order to test the feasibility of such an approach. While this experiment is conducted on a monolingual scenario, conclusions can still be drawn on the reliability of automatic interpretation of prosodic features.

6.4.1 Data

The database described in Sect. 5.2 consists of read speech which contains limited to no prosodic effects. Therefore experiments on prosodic properties have to be conducted on different data. As currently no suitably labeled non-native database is available, we chose to analyze native Japanese speech [Svojanovsky 04]. The data for this experiment was collected for the JST/CREST ESP (Japan Science & Technology Agency / Core Research for Evolutional Science and Technology) project [Campbell 00] and is a subset of 150 hours of Japanese everyday speech. The speaker is a Japanese female wearing a microphone and a Minidisc recorder whole day. Minidisc recordings undergo a lossy compression; acoustic quality issues are discussed in [Campbell 02]. The data consists of 2649 Japanese “un” backchannels. 482 of these utterances are labeled, the remaining 2167 are unlabeled. Table 6.11 illustrates the numeric labels assigned to each backchannel and its meaning as speech act.

6.4.2 Features

Classification and clustering of non-verbal utterances was conducted based on the features described in Sect. 2.3, namely duration (abbreviated as *dur*), power (*Pwr*), pitch (*F0*) and glottal characteristics.

For this experiment, we consider power relative to the rest of the utterance. Therefore we have four subtypes of power based features:

Pwrmean: the mean power
Pwrmin: the minimum power
Pwrmax: the maximum power
Pwrpos: the position of the maximum power relative to the duration of the examined word or speech segment

There are several features that we derive from pitch in this experiment:

F0mean: the mean pitch
F0min: the minimum pitch
F0max: the maximum pitch
F0pos: the position of the maximum pitch relative to the duration of the examined word or segment
F0vcd: the occurrence of voiced pitch relative to the duration of the examined word or segment
F0grad: the gradient from the position of the minimum and the position of the maximum of the pitch, relative to the duration

From the wide range of glottal features the following are extracted from the speech signal at the position with the highest energy:

Adduction quotient (H1-H2): the difference between the first and the second harmonic indicates a vowel. The value changes when the open quotient rises. *H1-H2* is a well-used indication of opening quotient or adduction quotient.

Spectral tilt (H1-A3): the amplitude of the third formant relative to the first harmonic *H1-A3* is an evidence for spectral tilt and displays the abruptness of the cut off of the airflow.

Table 6.12 Frequency of the label numbers in development set (dev) and test set (test)

Number	1	2	3	4	5	6	7	8
dev	189	23	10	7	2	1	7	3
test	188	23	9	7	3	0	8	2

6.4.3 Experimental Setup

The feature vectors of the data are normalized to values between 0.0 and 1.0. The training data is unlabeled and consists of 2147 utterances. The labeled data is divided into 2 parts, the development set and the test set. Table 6.12 shows the frequency of the labels and the arrangement to the sets.

In the experiments, algorithms such as nearest neighbor, furthest neighbor, or a combination of split&merge and k-Means cluster the training set into a specific number of clusters.

Bottom Up: This method starts with as many clusters as vectors in the set. Each vector belongs to its own cluster. The two nearest clusters are merged. This step continues until the number of clusters is reduced to a specific value.

Top Down: In this method all vectors are initially gathered in one big cluster. Then this cluster is split into two new clusters. Any vectors in this cluster are assigned to one of two new clusters which are represented by two vectors from this cluster. These two vectors can be the two furthest neighbors in the old cluster. Since the calculation of the furthest neighbors costs time, an improved algorithm termed as *fast split* is established in this project. Here the furthest vector to the average center is calculated. This vector is the first representative of the new cluster, the second one is the furthest neighbor to this vector.

k–Means: The *k–Means* algorithm classifies all data to a given number of clusters. The representatives of the clusters are recalculated and the clusters are emptied. Then the data is classified again. With each step, the shape of the clusters becomes clearer. The algorithm stops after a specific number of steps or a special criterion. The quality of the cluster strongly depends on the initial clusters. In this approach the initial clusters are represented by arbitrarily chosen vectors from the training data set. Three different initial vectors are tested in this project.

Split & Merge: This method is a combination of the *Bottom Up* and *Top Down* approach. The widest cluster is split into two new clusters until a certain number of clusters is produced. Then the nearest clusters are merged again. The number of clusters to produce changes with every single step.

Bottom Up + k–Means: In this method the clusters will be produced with the *Bottom Up* method explained above. But during the clustering process after a frequently number of merging steps, the *k–Means* algorithm based on the already existing clusters runs. Then the *Bottom Up* continues.

Split & Merge + k–Means: This is an extended version of *Split & Merge* algorithm. Every time the algorithm swaps from merge mode to split mode and from split mode to merge mode the *k–Means* algorithm runs to reshape the produced clusters.

The clustering algorithms are based on a distance measure between clusters. After preliminary experiments we chose the furthest neighbor distance. This method measures the distance between the furthest vectors from two clusters, which is the maximum distance of all vector pairs between both clusters.

As most of the available data is unlabeled, the experiment is setup as described in Fig. 6.7. First, the unlabeled training data set undergoes unsupervised clustering. By using the clusters to evaluate the development set, each cluster is assigned a label, resulting in a labeled cluster set. This labeled cluster set is the classifier on which the final recognition rate is calculated.

For an experimental setup, a clustering method and a weighting of features of the distance calculation is used. The first cycle of an experiment clustered the set into a number of clusters between 10 and 20. In a second cycle, the set is clustered into 50 clusters. Both results are evaluated and compared. A setup which indicates significant improvements by increasing the number of clusters is probably more useful to solve the existing problem than one without improvements. Obviously, these setups cluster the set into clusters which are more representative. Setups with few improvements scale the problem down. A 2-level classification will not solve the problem, because sub-clusters would not constitute significant better representatives.

6.4.4 Classes and Features

A visualization of the data and the experiments lead to the conclusion, that class 1 and 2, 4 and 5, as well as 7 and 8, are hardly separable; their behavior is too

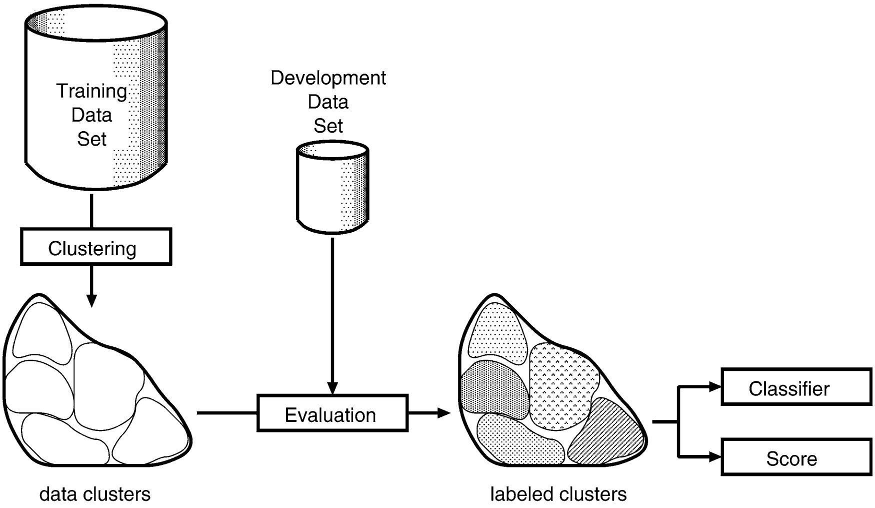


Fig. 6.7 Clustering algorithm utilizing the large unlabeled training set

similar. But due to the fact that their meaning is also similar, these classes can be merged. Figure 6.8 illustrates the distribution of the classes in the feature space, reduced to the dimensions *F0mean*, and *Dur*. It can be seen that Class 1+2 covers a large part of the space, while class 3 is scattered. Class 6 consists of only one vector. Class 7+8 overlap with 1+2, but are distinguishable from other classes.

Several preliminary experimental setups indicated that *F0mean, F0grad, Dur, H1-H2*, and *H1-A3* are more important than the power features. A higher weighting of these features improved the clustering and classification. This result covers the observations of Umeno [Umeno 03]. Disregarding the prosodic features *Pwrmax, Pwrmin*, and *Pwrmean* improved the performance, combined with giving high weights to *F0mean, F0grad, Dur, H1-H2*, and *H1-A3*.

6.4.5 Clustering Methods

The experiments proved, that neither a nearest neighbor nor a furthest neighbor clustering is sufficient for the existing problem. No weighting of the features achieved an average class recognition greater than 40%. A weighting showed no significant improvements. Many vectors on the edge of the training set are formed to clusters, which consist of just one or two vectors. These are not representative for the set. About 80% of the development set are assigned to one or two clusters in the center of the training set. The classification assigned no or just one vector to more than 60% of the formed clusters. Table 6.13 shows the average class recognition rate and the percentage of misclassified vectors of each clustering method.

Fig. 6.8 Visualization of non-verbal utterance classes in the labeled development set

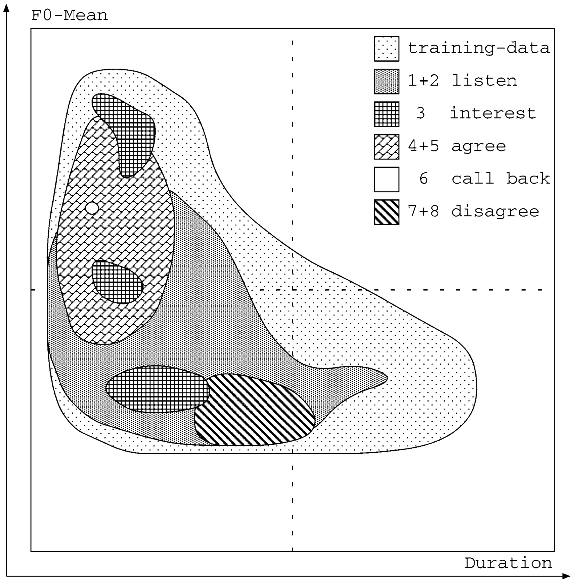


Table 6.13 Comparison of clustering methods: Average class recognition rate (recog) and misclassification (error)

	Not weighted		Weighted	
	Recog (%)	Error (%)	Recog (%)	Error (%)
Bottom up	60.9	36.0	53.2	43.6
Top down	42.1	56.4	59.3	34.9
S&M	60.3	53.1	57.1	44.8
K-Means	54.4	29.8	62.9	42.7
S&M + K-Means	61.3	50.2	68.6	39.2

A combination of K-Means and split&merge algorithms leads to the best results. After classification, the development set is distributed more uniformly over the clusters. 68.6% of the classes could be recognized using only 15 clusters, but still 39.2% of the vectors are classified incorrectly. A high number of vectors of class 1+2 are assigned to incorrect classes. After evaluation of these clusters, a further clustering setup ran on 5 of these clusters and created 5 to 10 sub-clusters. The weightings differed in these setups.

A 2-level classification achieves an average class recognition rate of 83.9% on the development set, while the misclassification of vectors is only 27.7%. The test set classification leads to the confusion matrix illustrated in Table 6.14. Classifying class 6 is impossible as there is only one instance in the data.

Table 6.14 Confusion matrix on test set using a weighted 2-level classification. Absolute number (n) and percentage of vectors in each class is given

Class	Classified as							
	1+2		3		4+5		7+8	
	n	%	n	%	n	%	n	%
1+2	148	70.1	26	12.3	13	6.2	24	11.4
3	4	44.4	4	44.4	1	11.1	0	0.0
4+5	3	30.0	0	0.0	7	70.0	0	0.0
7+8	2	20.0	0	0.0	0	0.0	8	80.0

6.4.6 Results

A 2-level classification achieves an average recognition rate of 66.1%. A combination of split&merge and k-Means clustering algorithms in both levels leads to this result. The power extrema showed to be unimportant for clustering, probably due to the differing recording conditions, while *F0mean, F0grad, Dur, H1-H2*, and *H1-A3* distinguish as most relevant features.

The first level of classification already achieves an average recognition rate of 64.5%. This rough estimate has a misclassification of 62.9%. The second level does not improve the average recognition rate significantly, but the misclassification shrinks to a value of 30.4%. In this step the estimated class is separated from other classes, mainly class 1+2. Class 4+5 distinguished in *H1-A3, F0-Grad*, and *F0max*, while class 7+8 distinguished in *Dur* and *F0max*. Class 3 causes problems, because the feature vectors are scattered over whole feature space. Because of the lack of data, a reasonable classification and testing of class 6 is impossible. Class 1 and 2, 4 and 5, as well as 7 and 8 are merged, because behavior and meaning is too similar to achieve a satisfying classification without being aware of the context.

6.5 Lessons Learned

From the experiments described in this chapter, we can conclude: The most straightforward approach to deal with non-native pronunciation variations is to generate phoneme confusion rules. This method is also very popular in literature and does improve recognition results, but only to a limited extend. Either rules are chosen based on heuristically determined thresholds, or an explosion of alternatives must be dealt with. Hitherto it seems more promising to model pronunciation variations statistically and implicitly in HMM models, as this type of model will be effective without thresholds. The increase of the number of possible pronunciation alternatives can be handled by applying the pronunciation network in the post-processing phase as rescoring function.

The ATR non-native speech database collected herein has proven to be a good basis for CALL research, with the proposed automatic pronunciation scoring features and algorithm being almost as reliable as human experts. We can also see that in special situations like a limited resource scenario with semi-continuous HMM acoustic models, local solutions as multilingual weighted codebooks can lead to good improvements.

Attempting the classification of non-verbal utterances remains a challenge. Although fillwords used to provide backchannel information during a dialog is possible to some extend, phoneme-level classification can be conducted with much higher reliability.

Chapter 7
Pronunciation HMMs

As the literature and experiments introduced in the preceding chapters showed, the performance for non-native speech recognition can be improved by considering the variations in pronunciation.

But we also saw that an explicit formulation of these variations e.g. as rules leads to a tradeoff between the number of variations considered and the dictionary or lattice size, which leads to confusions and again to a decrease in recognition accuracy. An unsolved problem about rules is also how to deal with insertions and deletions. A model is needed that handles pronunciation variations implicitly, and with sufficient context to cover insertions or deletions as well.

A further requirement for a promising approach of treating pronunciation variations is to take the statistical properties of these variations into account. Some variations are more frequent than others.

Finally, only data-driven approaches are powerful enough to be employed for a multitude of accents. Expert analysis of specific accents could lead to more precise models, but this seems to be impractical for real-world systems that have to deal with many native/non-native language pairs. We need a method that can extract pronunciation variations automatically from non-native training data without knowledge-based expert interference. Ideally, this method should also be able to allow for unseen pronunciation variations, as training data is always limited and some variations might not have been encountered during training.

These requirements can be covered by our proposed method [Gruhn 04c]: HMM-based statistical pronunciation models. The HMMs handle all variations implicitly and take the statistical properties of the variations into account. They can be derived and trained automatically from baseline pronunciation dictionaries and non-native training data. They even provide the possibility to allow unseen pronunciation variations with some minimum probability.

In this chapter, we explain the generation of pronunciation HMMs, the training approach and how they are applied to increase recognition rates for non-native speech.

R. E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*, Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_7,

7.1 Why HMMs?

There are many ways to model probabilistic events. Other than HMMs, for example neural networks and finite state automatons are common in the field of speech recognition.

Non-native speech data is generally in rather short supply, and we cannot hope for too many instances of individual words in training data. Some words even have to be modeled without training data. HMMs are more straightforward to initialize to a specific behavior than neural networks with hidden layers. But the main problem about neural networks is that they are discriminative models, which generally require more training data in order to avoid local effects impairing model behaviour. HMMs are generative models and are less error-prone in a scarce training data scenario. Finally, during application the desired result is a score of similarity rather than a classification result, which is more straightforward to extract from HMMs.

Finite state automatons (FSA) are another common generative modeling approach in speech recognition. The basic architecture of such a model would be similar to the modeling with HMMs, and it can be assumed that the pronunciation models could also have been examined with FSAs. HMMs were preferred because of convenient and training software being available for HMMs and the author's familiarity with it rather than any general theoretic reason.

7.2 Generation and Initialization

For each word in the vocabulary, one discrete untied HMM is generated.

The models are initialized on the phoneme sequence in some baseline pronunciation lexicon. The number of states for a word model is set to be the number of phonemes in the baseline pronunciation, plus enter and exit states.

Each state has a discrete probability distribution of all phonemes. The phoneme sequence(s) in the baseline dictionary are given a high probability and all other phonemes some low but non-zero value. Forward transition between all states is allowed, with initial transition probabilities favouring a path that hits each state once. The more states are skipped the lower is the transition probability.

Phoneme deletions are covered by state skip transitions, phoneme insertions are modeled by state self-loop transitions. The models may also be able to deal with spontaneous speech effects such as word fragments or restarts, which are basically consecutive phoneme deletions. But this has not been further analyzed in these experiments, as such effects do no occur in read speech.

Figure 7.1 shows as an example the HMM for the word "and" initialized from two pronunciation variants "ae n d" and "ax n d". The first state has output probabilities of almost 0.5 for each /ae/ and /ax/, and some small probability for all other phonemes (which are not shown in the figure). The output probabilities add

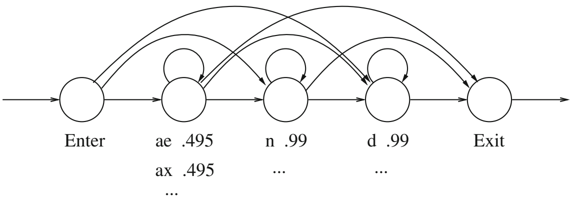


Fig. 7.1 An example discrete word HMM for the word “and”, initialized with two pronunciation variations for the first phoneme

up to one. As the second and third phonemes of that word are assumed to be identical in the two pronunciation alternatives of this example, they are assigned a probability of around 0.99 in the second and third state, respectively. Each state also has self-loops to allow insertions.

7.3 Training

As illustrated in Fig. 7.2, two levels of HMM-based recognition are involved in pronunciation HMM training:

- Acoustic level: phoneme recognition to generate the phoneme sequence S_i from the acoustic features O_i
- Phoneme label level: For training, the phoneme sequences S_i are considered as input. For all words, a discrete word HMM is trained on all instances of that word in the training data. The models are applied for rescoring, generating a pronunciation score given the observed phoneme sequence S_i and the word sequence.

The first step requires a standard HMM acoustic model, and preferably some phoneme bigram language model as phonotactic constraint. The continuous training speech data is segmented to word chunks based on time information generated by Viterbi alignment. Acoustic feature vectors are decoded to an 1-best sequence of phonemes.

The probability distribution as well as the transition probabilities are re-estimated on the phoneme sequences of the training data. For each word, all instances in the training data are collected and analyzed. The number of states of each word model remains unchanged.

Data sparseness is a common problem for automatically trained pronunciation modeling algorithms. In this approach, pronunciations for words that do appear sufficiently frequent in the training data, the pronunciations are generated in a data-driven manner.

The training of the pronunciation models takes place on the training data of every speaker of the regarding accent group, the test utterances are not included. The rather small number of speakers per accent group made a speaker-close setup necessary. In total, five models are trained. For rare words, the algorithm falls back to the baseform phoneme sequences from a given lexicon while still allowing

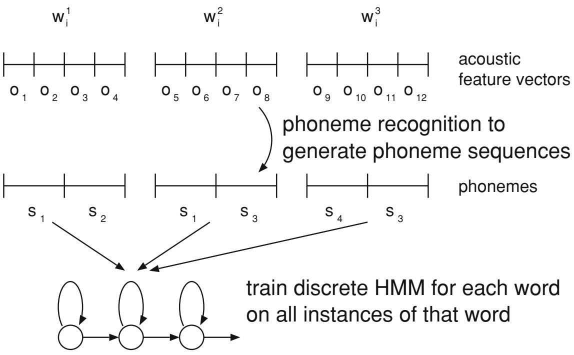

w_i^1
w_i^2
w_i^3
acoustic feature vectors
o_1 o_2 o_3 o_4 o_5 o_6 o_7 o_8 o_9 o_{10} o_{11} o_{12}
phoneme recognition to generate phoneme sequences
phonemes
s_1 s_2 s_1 s_3 s_4 s_3
train discrete HMM for each word on all instances of that word

Fig. 7.2 Two layers of processing are required to train pronunciation models: an acoustic level for phoneme recognition and the phoneme label level for word model training

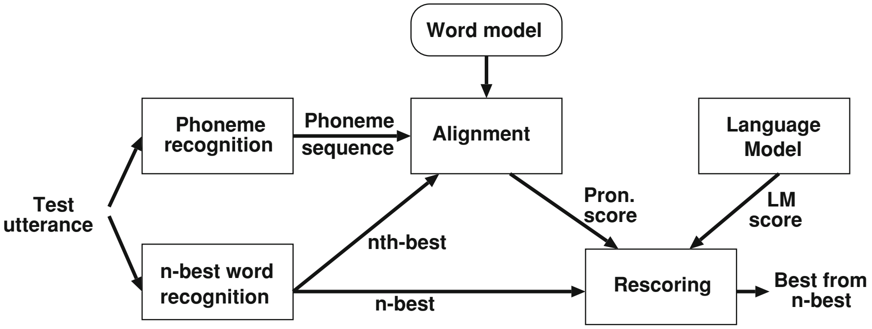


Fig. 7.3 Rescoring an N-best recognition result with word pronunciation models

unseen pronunciations with a low likelihood. This combination is expected to make it more robust than for example an application of phoneme confusion rules on a lexicon could be.

7.4 Application

Figure 7.3 shows the way pronunciation word models are applied by rescoring an N-best recognition result. On a non-native test utterance, both a 1-best phoneme recognition and a N-best (word-level) recognition steps are performed.

In standard Viterbi alignment, a speech signal is aligned to a reference text transcription using an acoustic model, with an acoustic score as a by-product Figure 7.4. In this approach, the time-aligned lattice is of no interest, although usually it is the main target of Viterbi alignment. Figure 7.5 gives a graphical explanation.

With the pronunciation HMMs as "acoustic model" and each N-best hypothesis as reference, a Viterbi alignment results in an "acoustic score", which is in fact the pronunciation score. Additionally to the pronunciation score calculated from the

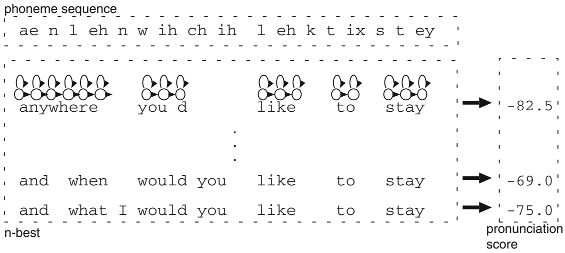

phoneme sequence
ae n l eh n w ih ch ih l eh k t ix s t ey
anywhere you d like to stay
-82.5
and when would you like to stay
-69.0
and what I would you like to stay
-75.0
n-best
pronunciation score

Fig. 7.4 For each n-best hypothesis of an utterance (*bottom three lines*), a pronunciation score is calulated relative to the phoneme sequence (*top line*). The correct result is "and when would you like to stay"

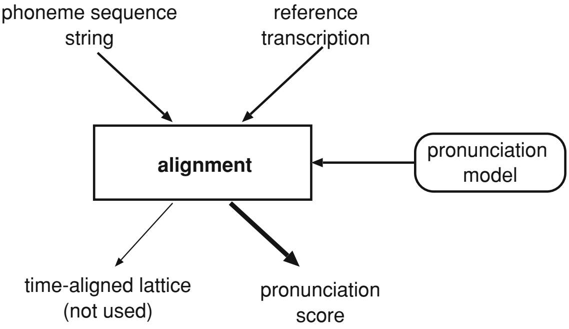


Fig. 7.5 The Viterbi alignment algorithm is used to calculate the pronunciation score

pronunciation HMMs, a word bigram model is applied to equate a language model score for each of the N best. Similar to (acoustic) speech recognition, the weighted combination of language model and pronunciation score turned out to perform best. The total score for each hypothesis i is calculated as $S_{total}(i) = S_{acoustic}(i) + \lambda S_{LM}(i)$. The hypothesis achieving the highest total combined score among the N-best is selected as correct.

Phonetic networks in general could also be applied directly in the decoder, for example by constructing parallel models [Minematsu 03] or incorporating into a Bayesian Network [Sakti 07]. But the high degree of freedom such models allow will cause additional recognition errors, as the proposed method allows *any* pronunciation with a minimum likelihood. Pronunciation networks are more reliable if applied in a separate post-processing step.

7.5 Experimental Setup

7.5.1 Data and Software

Additionally to the database presented in Chap. 5, the following standard databases have been part of the experimental setup:

Wall Street Journal (WSJ) Corpus The Wall Street Journal Corpus [Paul 92] was collected in two phases: the pilot project CSR-WSJ0 in 1991 and the main project CSR-WSJ1 from 1992 to 1993. The collection has been sponsored by the Advanced Research Projects Agency (ARPA) and the Linguistic Data Consortium (LDC) and has been carried out by MIT, Texas Instruments and SRI International. WSJ1 contains about 73 hours (approx. 78,000 utterances) of speech for training and 8 hours (approx. 8,200 utterances) of speech for testing purposes. The speakers of the training set are from North America and read texts from the Wall Street Journal newspaper. The data was recorded with a Sennheiser close-talking head-mounted microphone.

In this research all acoustic models are trained on the WSJ database. The selected training set consists of about 30,000 utterances of 200 speakers from WSJ1 and 7,200 utterances from WSJ0.

In order to evaluate the performance of the acoustic model built with the training data, the Hub2 test set was used. It comprises 20 utterances of ten native speakers each.

Furthermore, a phoneme bigram model was trained on the result of a forced alignment of the WSJ corpus.

TIMIT Corpus The TIMIT corpus [Fisher 87] was sponsored by the Defense Advanced Research Project Agency (DARPA) and set up by Texas Instruments (TI), MIT and SRI. The corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from eight major dialect regions of the U.S. These 10 sentences are from three different sets:

- SA: dialect sentences, which are meant to expose the dialectal variants of speakers within US English.
- SX: phonetically compact sentences, which are designed to provide a good coverage of pairs of phones.
- SI: phonetically diverse sentences, which are selected from the Brown Corpus and the Playwrights Dialog in order to add diversity in sentence types and phonetic contexts.

For none of the experiments described in this book the speech data of the TIMIT corpus has been used. Rather, sentence list of the TIMIT SX set was basis for the database collection described in Chap. 5.

SLDB The corpus used for English language modeling [Masataki 97] was chosen from *Speech and Language Data Base* (SLDB [Nakamura 96]) and the smaller LDB corpus, with a total of 2,072 conversations. The task is hotel and travel arrangement. In this data, each conversation consists of two conversation sides. Conversation sides are identified with labels: clerk, customer or interpreter. Here, clerk and customer are from native English speakers, interpreter denotes that it is a translated sentence from a Japanese-English dialog. Each dialog consists of typically 12 to 13 utterances. The average number of words per utterance is 11.

HTK is a standard toolkit for model training and speech recognition [Woodland 93]. All acoustic model training, language model training, speech recognition and forced alignment steps have been conducted with HTK version 3.2. HTK supports

Table 7.1 The English 43 model phoneme set in ARPAbet notation

AA AE AH AO AW AX AXR AY B CH D DH DX EH ER EY F G HH
IH IX IY JH K L M N NG OW P R S SH T TH UH UW V W Y Z SIL SP

discrete HMM models, therefore we could also train the pronunciation models with HTK tools.

7.5.2 Acoustic and Language Models

The acoustic models are trained on the native English speech data from the Wall Street Journal speech corpus described in Sect. 7.5.1.

The phoneme set consists of 43 phonemes, including two silence models, as listed in Table 7.1. Most research on the WSJ data bases on a similar phoneme set, which is also applied in this work. The notation is in capital letters and it contains the symbol *SIL* for silence. Following the usual HTK conventions, additionally the short-pause symbol *SP* is included.

Three acoustic models have been trained, a monophone, a right context biphone and a triphone model. The sizes are as follows:

- the monophone HMM model has 132 states and 16 mixture components per state,
- the biphone model 3000 states and 10 mixture components per state,
- the triphone model 9600 states and 12 mixture components per state.

All models have been trained with up to 16 mixtures, the numbers of mixtures listes above showed the best performance in phoneme recognition, as described in Sect. 7.6 To verify the general quality of the acoustic models and to ensure no mistake happened during training, the models were evaluated on the Hub2 task of the WSJ database. The word error rates of these models for the native English test set are 19.2%, 15.2% and 6.4%, respectively. These rates are comparable to standard results reported in the literature and at levels that can be expected for such a setup and task complexity. It can be concluded that there is no problem about the acoustic models and that they are suitable for further experiments.

With HTK, we created feature vectors from the speech files. The features are 12 MFCC coefficients, energy and the first and second derivatives, leading to a 39-dimensional feature vector. Cepstral mean normalization is applied.

All acoustic models are speaker independent models, we did not apply speaker adaptation methods such as MLLR. Adapting the acoustic model is of course a valid approach to increase recognition rates, but a different one, which was not the focus of this work. Pronunciation modeling and acoustic model adaptation address similar problems, both attempt to compensate the speakers pronunciation anomalies. Both approaches taken by themselves can improve recognition accuracy, and speaker adaptation gives more improvement than any type of pronunciation

Table 7.2 Phoneme accuracy in %, compared to a canonical transcription

	CH	FR	GER	IN	JP
Monophone	39.21	45.41	48.85	43.31	37.74
Biphone	29.54	37.87	41.15	33.84	29.24
Triphone	30.07	41.57	45.45	27.08	29.46

modeling. Doing pronunciation rescoring combined with adaptation does not help any further. The reason is that speaker adaptation is more specifically adapting to the speakers properties (not only his non-nativeness), something we do not attempt with pronunciation modeling. In our approach we are targeting an accent group rather than a specific person. In general, speaker adaptation methods can be applied for the whole group, but the point is that speaker adaptation makes the acoustic model speaker dependent, and every time there is a new speaker, retraining is necessary. With the statistical lexicon, the AM is still speaker independent.

A bigram language model was trained on the SLDB travel and hotel reservation domain database. The model is a simple word bigram language model without classes, even though the database contains part-of-speech tags.

Two types of dictionaries have been the base of both pronunciation HMM creation and N-best recognition, a LVCSR dictionary with 8875 entries for 7311 words is used in the main experiment. The 7.3k words fully cover a general travel scenario.

Some experiments that focus on the data collected in the first session consisting of a group of Japanese speakers of English were conducted with a more specialized hotel reservation dialogue dictionary of 6650 entries for 2712 words.

The test set contains 344 words in two dialogs with 23 utterances in total. In the mostly used 7,300 word dictionary case, the perplexity of the language model was 32.

7.6 Phoneme Recognition

As a data-driven approach, the pronunciation modeling method proposed in this book includes a phoneme recognition step. For native speakers, context-dependent acoustic models achieve higher accuracy than monophone models. To examine the impact of acoustic model context for non-native speakers, phoneme recognition was performed on full utterances with a monophone, (right-context) biphone and triphone model.

Table 7.2 shows the phoneme accuracy for monophone, biphone and triphone models on the non-native data. Acoustic models were trained with up to 16 mixture components for each context size; in Table 7.2, only the best performing number of mixtures (as listed above) is shown.

A phoneme bigram model trained on the result of a forced alignment of native speech (WSJ) provided some phonotactic constraint.

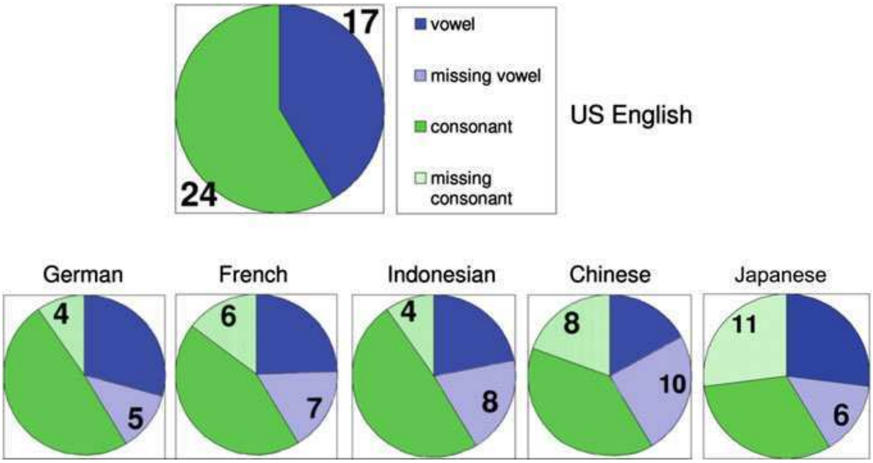


Fig. 7.6 Phoneme coverage of US English and selected other languages, for consonants and vowels

The reference for evaluation is also a canonical transcription generated automatically from a baseline native speaker oriented lexicon by forced alignment. This forced alignment can only include those variations included in the baseline pronunciation dictionary. It relies of the speakers saying exactly what they are supposed to say. If a correct phoneme transcription were available, higher numbers could be expected. For context dependent units, we consider only the center phonemes for calculating recognition accuracy.

The monophone model performs best for all speaker groups. Although the tendencies in recognition rates are not exactly the same for all data points, there clearly is an overall trend: The less context the models include, the better is the phoneme accuracy. We can conclude that the phonetic context for native English speakers is considerably different to non-native speakers.

The phoneme recognition rates are somewhat lower than what is usually seen for native speech and not at the same level for all speaker groups. Although there may be random differences in the educational background of the speakers, a more likely reason can be found when comparing the American English and regarding native IPA phoneme sets: German has at least 28, Indonesian 26, French 25, Mandarin Chinese 21 and Japanese 19 phones in common with American English. This makes it intrinsically more difficult for Chinese and Japanese speakers to achieve as high phoneme (and word) accuracies as the other speaker groups. Figure 7.6 visualizes the missing phonemes, divided by consonants and vowels.

The phoneme recognition rates on the non-native data may appear somewhat low, but they are at levels that have to be expected for non-native speakers Table 7.3. For comparison, Table 7.4 shows phoneme recognition results for the monophone and triphone acoustic models for the WSJ Hub2 test set. Reference was a forced alignment result. In the native case, triphones perform much better

Table 7.3 Distribution of errors for the monophone model (relative share of total errors)

	CH	FR	GER	IN	JP
Del	17.62	21.84	23.94	20.31	18.36
Ins	19.52	14.54	14.53	14.90	18.24
Sub	62.85	63.60	61.51	64.77	63.38

Table 7.4 Phoneme recognition performance in % for the native English WSJ database Hub2 task

Acoustic Model	Correct	Accuracy
Monophone	36.41	32.71
Triphone	70.97	67.31

than monophones and a phoneme recognition accuracy of 67.31% could be achieved.

7.7 Rescoring with Pronunciation HMMs

7.7.1 Word HMM Initialization

The discrete probability distribution for each state is initialized depending on the baseform phoneme sequence(s) as given in the lexicon. The correct phoneme has a probability of 0.99. If more than one pronunciation variant is included in the lexicon, the variations all have the same probability, totaling 0.99. The possibility that some pronunciation variants are more frequent than others can not be taken into account at this stage. All other phonemes are assigned some very small but non-zero probability.

The transition probabilities depend on the number of succeeding phonemes in the baseline lexicon. The probability to skip k phonemes is initialized to 0.05^k. Insertions are allowed with a chance of 0.05. The transition to the next state therefore has a probability of slightly below 0.9, so that the probabilities add up to 1. In our experiments, we allowed up to three skips in the initial models. Those numbers are chosen heuristically and showed to perform acceptably well. As the probabilities are modified in the training process, they are mainly relevant for words for which no training samples are available.

7.7.2 Training of Word HMMs

For the training of the word models, the non-native training data set is segmented into single words based on time information acquired by Viterbi alignment. On these word chunks, phoneme recognition is performed, as explained in Sect. 7.3 The phoneme recognition result is converted into discrete features, with one feature file for each word of the training data.

Table 7.5 Word error rates in % for non-native speech recognition without and with pronunciation rescoring

	CH	FR	GER	IN	JP	avg
Baseline	51.23	37.93	31.77	40.48	56.92	45.88
Rescoring	45.12	34.80	29.88	38.31	52.36	42.14
Rel. impr.	11.93	8.25	5.94	5.36	8.01	8.15

7.7.3 Rescoring with Word HMMs

The HMM pronunciation models are applied in the form of rescoring the N-best decoding result as described in Sect. 7.4. Table 7.5 shows word error rates with and without pronunciation rescoring. The numbers have been calculated for the complete database and the large vocabulary setup.

There are several aspects about the rescoring algorithm that require closer examination in order to find a deeper understanding of the benefits of the proposed statistical pronunciation models. The following experiments we conducted on the smaller setup of 11 Japanese speakers and with the more specialized 2700 word dictionary.

7.7.3.1 Language Model Weight

One issue is the language model scale factor. Rule-based approaches usually find that if too many pronunciation variants are added to a dictionary, the recognition rate drops again because additional confusions occur. Some papers (e.g. [Kim 07]) try to conceal this effect by factoring in a single-digit perplexity language model. In such a setup, it is advantageous to allow as many words as possible to be permitted in the decoder: The language model will, if scaled high enough, remove all the nonsense confusions. Such a setup is not valid for real-world systems, as in real world speech recognition systems usually the speaker has more freedom to make utterances, and if the dialog structure actually is that strict or the vocabulary size that low, there is little need for pronunciation modeling. With the proposed statistical pronunciation models, any pronunciation is basically possible as all phoneme sequences are given a non-zero likelihood. This makes an analysis of the influence of the language model necessary.

The language model score is added to the pronunciation score with a weighting factor λ. Figure 7.7 shows the word error rate dependent on λ. The baseline performance (horizontal line) of 32.54% word error rate can be improved to 29.04%. The correct choice of the language model score weight is important, in this experiment a factor of 5 was the optimum. It can be seen clearly that the performance is best if the weighting of the language model score is neither too low nor inappropriately high. This means that there is a clear positive effect from pronunciation modeling and that factoring in a language model does help. But it is not the only relevant contributor, as overweighting the language model causes the error rate to rise again.

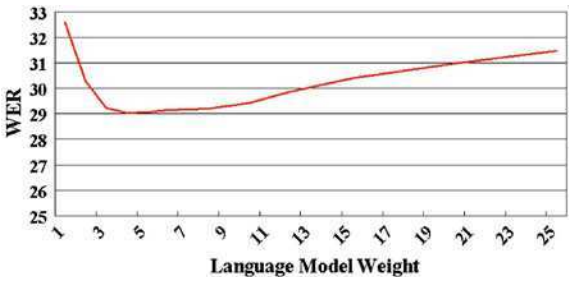


Fig. 7.7 Word error rate for rescoring of N-best based on pronunciation score combined with weighted language model scores

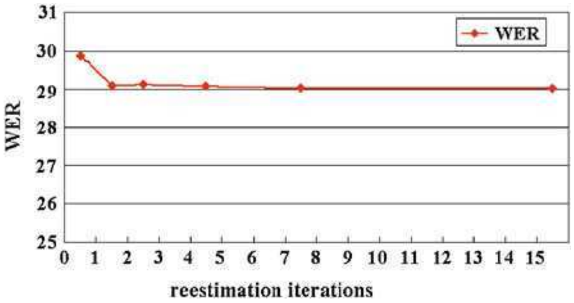


Fig. 7.8 Word error rate for rescoring of N-best based on pronunciation score combined with weighted language model scores

7.7.3.2 Training Iterations

The pronunciation HMMs are initialized from the baseline pronunciation dictionary, then several re-estimation iterations modify the probabilities. The effect of these training steps can be seen in Fig. 7.8, showing the results of experiments on the development set. Most improvement can be gained with the initial models already, from 32.54 to 29.88% WER. The first training iteration reduces the WER to 29.11%, further iterations bring only minor improvement. Limited coverage of the test data due to small training data may be the reason why the effect of increased training is limited.

7.7.3.3 Factoring in the Acoustic Score

In the previous experiments, the pronunciation score was combined with a weighted language model score. Rescoring only on the basis of the pronunciation score did improve the word error rate. But the pronunciation information alone did not perform as well as when language model information was added.

One possible extension would be to take the acoustic score into account additionally, or instead of the language model score. The acoustic score for each of the hypotheses is calculated at the N-best recognition step and therefore the additional consideration does not cause any extra computation cost. The acoustic score can be weighted relative to the pronunciation (and language model) scores.

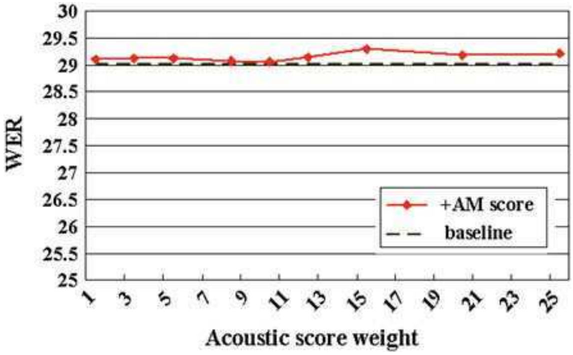


Fig. 7.9 Considering the acoustic score additionally to pronunciation and language model score does not lead to WER reduction

But it turns out that considering the acoustic score for rescoring does not lead to any improvement. The results of an experiment conducted on the smaller set of Japanese speakers is shown in Fig. 7.9. The baseline system (horizontal dashed line) considers only pronunciation and language model score, the language model weight is set to 5. Independent from the acoustic score weight, the baseline system always performs slightly better than the system considering the acoustic score. Results including the acoustic score at various scaling factors basically are all the same without showing any statistically significant tendency.

7.7.3.4 Data Coverage

For acoustic models, phonemes are usually chosen as basic unit rather than words. The reason is that for such small units, the number of training items is much higher than for large units. For pronunciation modeling, phonemes are not the optimal unit because of the strong context or even word dependency of pronunciation errors. The drawback of large units is that not all words can be trained: Only 15.3% of the 7312 words are in the training data and can serve to reestimate the HMM parameters, the rest of the word models remains in the initial state.

A level between words and phonemes are syllables. The words in the pronunciation dictionary can be split into syllables automatically with syllabification software such as [Fisher 96], resulting in 3422 distinct syllables. While the dictionary size decreases, the number of units in the test data actually increases, from 141 words to 182 syllables. While this makes syllables unlikely to be a better unit for pronunciation modeling, other sub-word units such as syllable compounds might be an interesting target for future experiments.

All in all, the results show that pronunciation HMMs as statistical lexicon are an effective way to increase performance for non-native speech recognition. An average recognition rate increase of relative 8.2% was achieved. The improvement has been proven to be causal by pronunciation rescoring. The remaining gap to native performance could be narrowed by increased training data coverage. To achieve this, it is necessary to either increase the amount of training data even further or choose units other than words for modeling with HMMs.

Chapter 8
Outlook

In this research, we proposed HMM based statistical pronunciation modeling for non-native speech recognition. This method is data-driven, and has shown to be applicable to any given accent type without need for expert knowledge about the specific accent.

Motivated from patterns visible in a phoneme confusion matrix, rules seem capable of representing pronunciation variations of accented speech. Literature shows rules are a popular form to convey and apply pronunciation information to the dictionary of a speech recognition system. But as we have shown, there are limitations to the rule-based approach: Less frequent variations have to be disregarded, as they otherwise cause additional confusion.

Pronunciation HMMs are a soft form of representation that allows to retain more of the information given in the training data than explicit representations such as rules which require hard decisions. Rules also require to search for thresholds in order to balance additional pronunciation variations in a dictionary with the number of additional confusions caused.

In our proposed statistical dictionary, the pronunciation variation information is represented implicitly in the form of HMMs. The models are applied by calculating a pronunciation score with the Viterbi alignment technique and rescoring the n-best hypothesis list. Handling a model that allows a multitude of possible pronunciations with different likelihoods through rescoring has shown to be robust against confusions, as in no experiment the rescoring decreased performance.

There are several databases in existence, but most of them are very small, of unsuitable acoustic condition or unavailable to the public. We therefore collected a database of non-native English speech as part of this work. It includes 96 speakers of five accent groups (Mandarin Chinese, French, German, Indonesian, Japanese). With over 22 h of speech it is one of the largest existing non-native speech databases.

The data consists of read speech, such as hotel reservation dialogs, number sequences and phonetically rich sentences. Each speaker produced a uniform set of

R. E. Gruhn et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing*, Signals and Communication Technology, DOI: 10.1007/978-3-642-19586-0_8,

146 utterances. The data collection was conducted with special care to avoid speaker anxiousness.

Pronunciation skill ratings on sentence and word level by human experts make the database applicable for research in the field of computer assisted language learning. For such ratings, the learning process of the experts during the scoring could be a problem: humans might chance their rating behavior over time. We aimed at avoiding such effects by calibrating the experts with a uniform initial sentence set and presenting the target utterances in random order to several raters.

The database yields a higher number of male speakers, however for the purpose of non-native accent analysis, this gender imbalance did not show any impacts on the results. The data is available for a fee at ATR, a well-known collector and distributor of speech databases in Japan.

For all experiments with non-native speakers, monophone acoustic models performed better than context-dependent ones. The coarticulation effects significantly differ between native and non-native speakers, and the more specialized context-dependent models do not match non-native speech sufficiently well.

Pronunciation scoring improved the word error rate in average from 45.88 to 42.15%, a relative error rate reduction of 8.15%. This demonstrates the effectiveness of the proposed approach in a non-native speech recognition scenario. The highest improvement was achieved for the Chinese accented speakers, with a relative 11.93% gain. The tendency of improvement was the same for all accent groups. Differences in the baseline performance can be explained with the smaller overlap of the phoneme sets especially of the Asian languages with the American English set. The German speakers obtained the highest baseline performance, due to the high number of English phonemes also existing in the German language.

The rescoring procedure provides best results if the pronunciation score is considered in combination with a weighted language model score calculated from a word bigram. Overweighting the language model reduced performance again, proving that the contribution from the pronunciation model is important. Factoring in the acoustic score on the other hand turned out to be hurting rather than helping during the rescoring process.

The pronunciation HMMs in the initialized state already lead to some improvement through the pronunciational relaxation provided by the models. This supports the initialization concept. Training the models on the results of a phoneme recognition further increases their modeling accuracy as expected. As our pronunciation model is applied as rescoring, we can take advantage of the multitude of allowed pronunciation variants without causing confusion. This would be the case with all other dictionary modification approaches.

A possible extension to our pronunciation modeling approach seems to lie in the combination with acoustic model adaptation. Pronunciation modeling and acoustic model adaptation address similar problems. They attempt to compensate the speakers pronunciation anomalies. Both approaches taken by themselves may improve recognition accuracy, with MAP speaker adaptation performing better compared to pronunciation modeling. Combining pronunciation rescoring with acoustic model adaptation does not seem to help any further. The reason is that

MAP speaker adaptation is more specifically adapting to the speakers properties (not only his non-nativeness). This was not attempted with the pronunciation modeling research reported here. We are targeting a general accent group rather than a specific speaker, leaving the acoustic model speaker independent.

Another remaining problem is the data coverage. From the available speech, two hotel reservation dialogs were chosen as test set randomly. Only a rather limited share of the vocabulary in the test set could be trained on the given data. As shown in this work, trained models can be expected to perform better than the models directly generated from a pronunciation dictionary. Many word models remain in the initial state with default transition and output probabilities. In a large vocabulary scenario, this gap can never be fully avoided. Modeling at the syllable level rather than at the word level does not seem to help. The coverage may be increased by applying additional non-native speech databases. However, the experiments have also shown that even unseen pronunciation variations are covered by the baseline models. This represents a clear advantage over other non-native speech pronunciation modeling methods.

For each accent group, a separate accent model is trained. In applications where the accent of the speaker is unknown, automatic classification of the accent will be necessary. Applying such an accent classifier is also left as future work.

While we have examined various measures for scoring the pronunciation of non-native speakers, we have not attempted to apply the pronunciation HMMs for this purpose. Training a native English pronunciation model and comparing the score of the native model with the accent models would be a possible experiment. The comparison would not be helpful for scoring though, as we could only measure similarity to the given accents. It is imaginable that such a comparison might be useful for accent classification.

It may also be helpful to initialize the transition probabilities in the pronunciation models based on an examination of typical insertion and deletion error frequencies rather than uniform. The expected gain would be small though, as the contribution of the transitions to the total score is comparatively low, and easy applicability of our approach would suffer somewhat. Still, further improving the initial models would be a possible continuation of this research.

We evaluated the (monophone) acoustic models on the native English Hub5 task and achieved a 19.2% error rate. For the hotel reservation task of the ATR non-native speech database, a similar performance could be expected for native speakers. Even with pronunciation rescoring, the error rates for non-natives are still below this level. Example for German speakers we reduced a 31.8% word error rate to 29.9%. Such an improvement of absolute 2% is helpful, but still less than the "gap to nativeness" of around 12% additional word error. Therefore even with this work the problem of non-native speech recognition can not be seen as summarily solved yet. Further improvement could be found in pronunciation models that take skill into account additionally to accent, but for such a specific model, training data is unlikely to ever be sufficient.

Real applications will have to deal with effects that cannot be handled on pronunciation level, such as the speaker completely omitting words. In such cases, only a dedicated language model could compensate. A further real-world problem are in-word pauses caused by hesitations of non-native speakers who have to rethink how to continue a sentence. The proposed pronunciation models could deal with this, as inserting silence within words can be allowed. But this effect is not seen in the test data, as our recording conditions of read speech prevent it.

Non-native speech recognition is necessary for applications such as speech-to-speech translation, car navigation, automatic telephone call processing and automatic travel assistance. The latter two applications are typically on server systems where computation is not a central issue. The first two systems run typically on low-resource hardware. The proposed method adds phoneme recognition and a score calculation with the Viterbi algorithm to the standard speech recognizer setup. But as both additions are lightweight in terms of calculation cost and memory consumption, they can be implemented even in low-resource systems such as embedded hardware or hand-held computers.

Therefore we can conclude that we have proposed a method that is effective in supporting non-native speech recognition and that is applicable in real-world systems. All in all, we achieved an average word error rate reduction of 8.15%, showing the effectiveness of the proposed method.

Appendix A: Hotel Reservation Dialog

This chapter gives an example for a hotel reservation dialog side as used in the data collection. The dialog shown here is TAC22012 as it is given in the transcription files. One side is a Japanese traveler (not listed below), the other side an English speaking hotel clerk.

good afternoon new york city hotel can i help you
okay for how many people would that be

okay and could i have your name please

okay mr. tanaka and when would you like the room for

okay so that 's one night for one adult and two children is that correct

okay uhmm we do have one twin room available on the fifteenth

and we could put a cot in the room for your other child

that 's okay the cots are quite big even adults can sleep on them

okay a twin room is a hundred and thirty dollars plus service charge and tax

i 'm sorry uhmm all the doubles are booked and so are all the singles which probably wouldn't be appropriate anyway

okay sure and would you like breakfast sir

sure we have the continental breakfast which is ten dollars and an english breakfast which is twelve dollars

so ahh will that be one continental breakfast or would you like it for the children also okay no problem and will you be paying by cash or charge sir

okay could i have your number in that case please

okay thank you and when does it expire

thank you and could i have your address and telephone number please mr. tanaka

so that ’s two six seven five two one zero three eight seven room five o seven washington hotel

okay mr. tanaka so my name is mary phillips and i ’ll be seeing you soon bye now

Appendix B: Confusion Matrices

This chapter shows phoneme confusion matrices for the five large accent groups described in Chap. 5. The images are comparing a phoneme recognition result obtained by recognition with the monophone acoustic model with a reference phoneme transcription obtained by forced alignment. The darker a box is, the more frequently the regarding confusion occured. The phoneme recognition result represents which sounds the speaker actually produced. The diagonal shows all cases of correct pronunciation, all other entries are mispronunciations. This simple model is of course somewhat distorted by random recognition errors, still the graphs show clearly some typical error patterns. The *x*-axis shows the recognition result, the *y*-axis the correct phoneme (Figs. B.1, B.2, B.3).

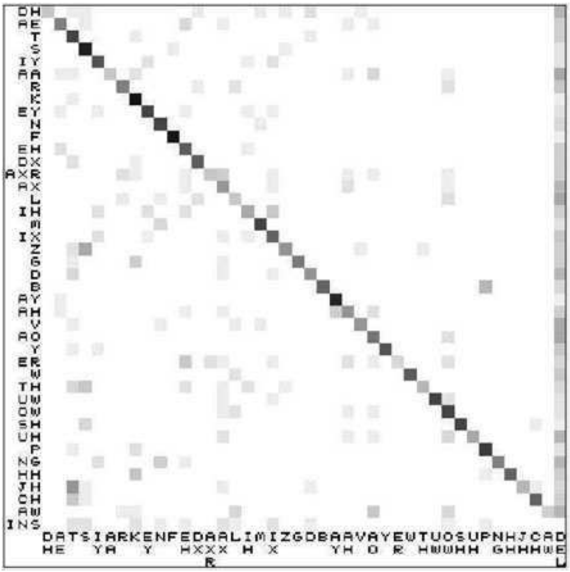


Fig. B.1 Phoneme confusion matrix for Mandarin Chinese speakers of English

Fig. B.2 Phoneme confusion matrix for French speakers of English

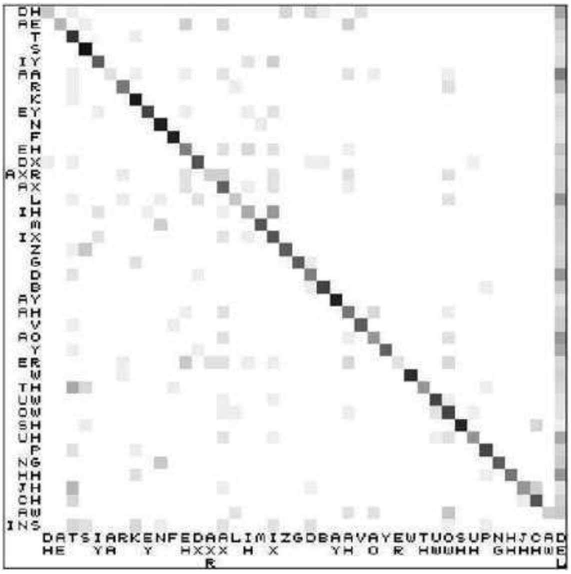
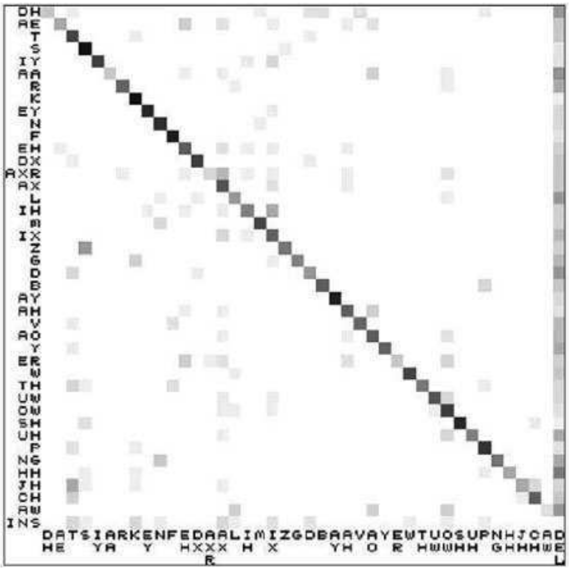


Fig. B.3 Phoneme confusion matrix for German speakers of English



The non-native speech database contains also three native speakers of English, one Australian, one British and one US-American (Figs. B.4, B.5). The data of three speakers of rather different types of English is not be the ideal basis to calculate a confusion matrix from, and only three speakers may not be enough to derive a pattern. But for the purpose of illustration we calculated a phoneme confusion matrix for native English as well, which is found in Fig. B.6.

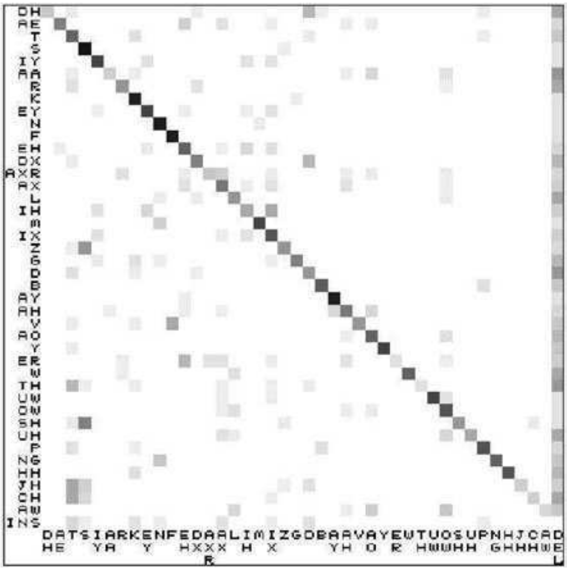


Fig. B.4 Phoneme confusion matrix for Indonesian speakers of English

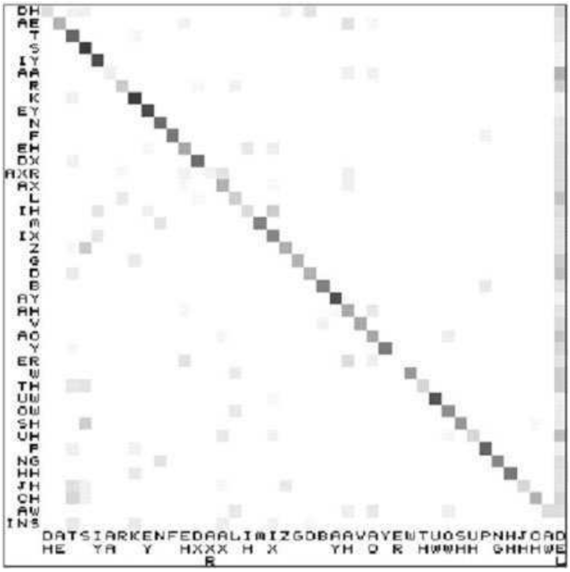


Fig. B.5 Phoneme confusion matrix for Japanese speakers of English

Substracting the confusion values of the native speakers from one of the accents' confusion matrix and calculating the absolute values matix yields an image visualizing the differences between native and non-native pronunciation and recognition properties. An example for such an image is Fig. B.7, where we compare Japanese English to a native English matrix. The darker, the greater the difference.

Fig. B.6 Phoneme confusion matrix for three native speakers of English

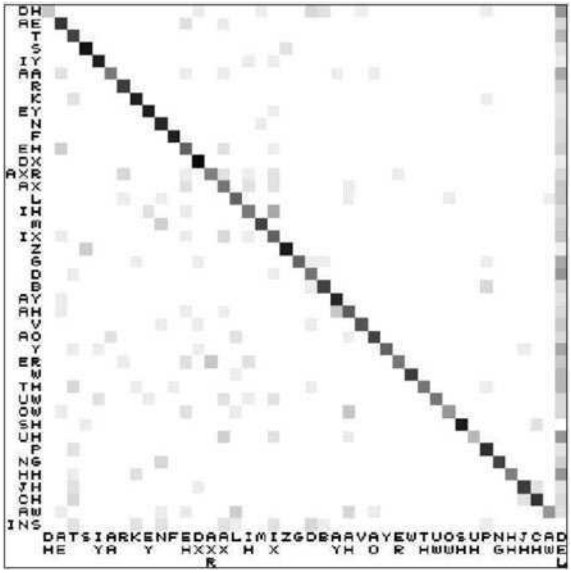
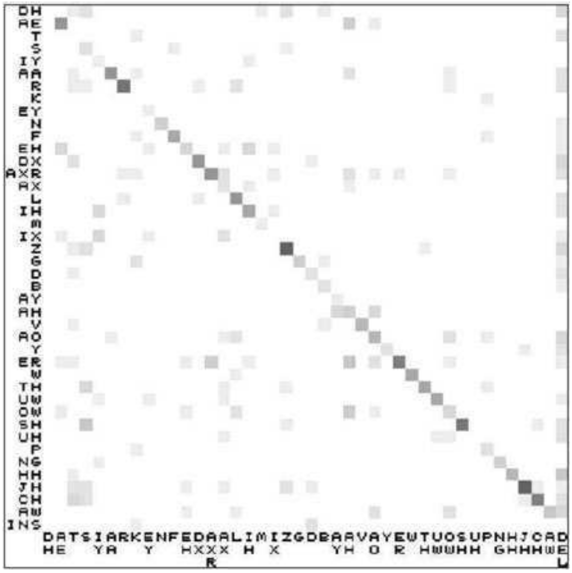


Fig. B.7 Difference matrix comparing native speakers of English with Japanese accented English



Appendix C: Speaker Information

This chapter explains specific details about the speakers. Each speaker is identified with a four-symbol ID number, consisting of the letter M or F identifying the gender followed by a running three-digit number.

For each speaker, the average human rating on utterance level, the number of mispronounced words (MisW), the ratio of mispronounced words (MisR), the phoneme recognition accuracy (PA) with a monophone acoustic model, age and first language are shown. For some Japanese speakers, the age information is not available. The table is sorted by the individual pronunciation skill ratings as calculated in [Cincarek 04a].

Speaker	Rating	MisW	MisR	PA	Age	Native
F018	1.03	7	0.02	21.27	-	Japanese
M076	1.30	29	0.07	26.48	43	German
M036	1.40	21	0.05	22.40	30	German
M052	1.43	17	0.04	38.35	36	German
M078	1.43	18	0.05	25.00	35	German
F022	1.60	27	0.07	18.43	45	Japanese
M001	1.65	33	0.08	33.99	39	German
M055	1.67	66	0.17	25.88	52	Chinese
M054	1.75	54	0.14	28.88	28	German
M040	1.77	41	0.10	11.56	21	Chinese
M051	1.77	62	0.16	32.90	26	German
M071	1.77	25	0.06	16.67	39	German
F026	1.80	56	0.14	25.39	28	Chinese
M042	1.82	52	0.13	31.59	33	Hungarian
M033	1.83	47	0.12	35.26	30	Indonesian
M056	1.83	59	0.15	28.95	35	German

(continued)

(continued)

Speaker	Rating	MisW	MisR	PA	Age	Native
M014	1.87	58	0.15	12.47	35	Japanese
F021	1.90	66	0.17	30.08	40	Korean
M044	1.90	71	0.18	17.50	25	French
M066	1.93	72	0.18	24.20	30	French
M026	1.95	71	0.18	32.33	42	French
M072	1.97	62	0.16	18.27	26	French
M032	2.00	49	0.12	14.16	32	Spanish
F009	2.02	77	0.19	22.52	-	Japanese
F025	2.03	41	0.10	22.55	35	Portuguese
F012	2.07	72	0.18	2.30	-	Japanese
M039	2.07	37	0.09	9.37	33	Sinhalese
M061	2.10	58	0.15	26.87	25	Indonesian
M073	2.10	37	0.09	28.25	27	French
F023	2.12	64	0.16	22.34	26	Japanese
F024	2.15	65	0.16	21.20	25	French
M045	2.17	64	0.16	30.95	23	French
M050	2.17	45	0.11	26.67	29	Indonesian
M010	2.20	85	0.22	22.68	25	German
M024	2.20	47	0.12	15.45	21	French
M034	2.20	86	0.22	26.67	23	German
F014	2.23	78	0.20	14.66	-	Japanese
M006	2.23	64	0.16	29.73	26	German
M043	2.25	74	0.19	20.97	24	French
M085	2.25	68	0.17	29.40	28	French
F010	2.27	63	0.16	15.21	-	Japanese
M021	2.30	72	0.18	23.69	26	German
M080	2.30	87	0.22	12.87	24	French
M022	2.38	90	0.23	1.50	42	Bulgarian
M059	2.38	62	0.16	6.66	43	Indonesian
M037	2.43	71	0.18	13.17	25	French
F019	2.47	62	0.16	18.45	26	Indonesian
M023	2.47	109	0.28	10.35	24	French
M030	2.47	82	0.21	12.14	35	Chinese
M077	2.47	103	0.26	11.32	30	French
M092	2.47	47	0.12	17.76	26	German
F008	2.50	83	0.21	18.42	-	Japanese
M016	2.53	60	0.15	-4.13	37	Japanese
F013	2.57	65	0.16	8.67	-	Japanese
M089	2.60	101	0.26	14.88	25	Indonesian
M060	2.62	97	0.25	13.59	24	Japanese
M035	2.65	121	0.31	19.51	23	French
M027	2.67	101	0.26	13.08	31	Hindi
M075	2.67	85	0.22	20.99	29	Indonesian
M068	2.70	95	0.24	0.41	32	Indonesian
M082	2.70	121	0.31	23.84	22	Japanese

(continued)

(continued)

Speaker	Rating	MisW	MisR	PA	Age	Native
M031	2.73	96	0.24	8.28	38	Indonesian
M058	2.73	88	0.22	2.57	39	Japanese
M063	2.73	93	0.24	23.04	36	Indonesian
M065	2.73	92	0.23	20.75	22	Japanese
M084	2.73	56	0.14	-13.75	25	Chinese
M074	2.75	107	0.27	6.45	25	Chinese
M011	2.77	78	0.20	6.78	41	Japanese
M012	2.83	100	0.25	5.42	29	Japanese
M067	2.83	96	0.24	4.20	25	Chinese
M087	2.83	88	0.22	11.48	24	Indonesian
M046	2.87	74	0.19	17.68	31	Indonesian
M053	2.87	79	0.20	7.33	40	Chinese
M029	2.88	121	0.31	18.24	22	French
M069	2.90	108	0.27	0.34	24	Japanese
M070	2.95	125	0.32	-2.98	22	Japanese
M041	2.98	124	0.31	16.62	30	Indonesian
F011	3.00	132	0.33	12.59	-	Japanese
M028	3.00	109	0.28	16.98	33	Chinese
M090	3.00	110	0.28	10.26	30	Indonesian
M083	3.05	131	0.33	4.68	24	Japanese
M038	3.10	108	0.27	15.92	30	Indonesian
M015	3.12	100	0.25	8.82	29	Japanese
M057	3.15	120	0.30	12.27	26	Chinese
M064	3.23	117	0.30	-3.25	21	Japanese
M013	3.30	76	0.19	-21.71	22	Japanese
F020	3.33	166	0.42	7.67	30	Chinese
M049	3.47	132	0.33	-2.99	33	Chinese
M086	3.58	124	0.31	8.63	28	Indonesian
M062	3.67	148	0.37	8.60	31	Chinese
M047	3.70	128	0.32	10.93	37	Chinese
M093	3.73	167	0.42	-17.39	31	Japanese
M088	3.93	181	0.46	3.45	27	Chinese
M025	3.97	216	0.55	1.09	37	Chinese
M081	4.07	118	0.30	-28.39	28	Chinese
M091	4.27	137	0.35	-4.61	31	Chinese

Appendix D: Human Evaluation

This chapter gives further information about the human expert ratings. The rating instructions are followed by the personal background of the evaluators and the list of speakers assigned to each rater. Finally screenshots of the evaluation browser interface are shown.

Evaluation Instructions

- The aim of this experiment is to obtain an assessment of proficiency of non-native speakers in terms of pronunciation and fluency.
- First we would like you to listen to 22 utterances. Each sentence is uttered by a different speaker. Please assign a level of proficiency to each utterance. Each level of proficiency should be selected at least once. Level 1 (green) should be used for maximum proficiency, Level 5 (red) for minimum proficiency present among speakers. The purpose of this step is to give a feeling of how to use the grading scale.
- Then you will start evaluation for a subset of 100 non-native speakers. There are 48 different utterances per speaker. They are presented in random order. You can listen twice to an utterance if it seems necessary to you. First we would like to ask you to mark any mispronounced words. Please tolerate any pronunciation variations which may exist in standard British or standard American English. Please consider only phonetic errors and ignore wrong lexical stress.
- There may be some words a speaker has (completely) misread. Please also mark these words as mispronounced.
- Next, we would like you to select a level of overall proficiency for the utterance by considering both pronunciation and fluency. Badly pronounced words, misread words, strong non-native accent, long pauses between words, stuttering, etc. should have an influence on your assessment. Please ignore sentence intonation for your evaluation.

- If there is an utterances which was not playable, please skip the sentence by selecting *NONE*, any proficiency level and clicking on the submit button. Please write down the number of any sentence for which the utterance was not playable on this instruction sheet.

Evaluator Information

Table D.1 from [Cincarek 04a] gives information about the human evaluators, who marked mispronounced words and assigned ratings to each utterance of the non-native speakers. The evaluators 3 and 16 grew up and went to school in Canada, all other evaluators in the United States.

Table D.2 from [Cincarek 04a] shows the utterances of which speakers were assigned to which rater for evaluation. All evaluators processed 1152 utterances (24 speakers times 48 sentences).

Rating Interface

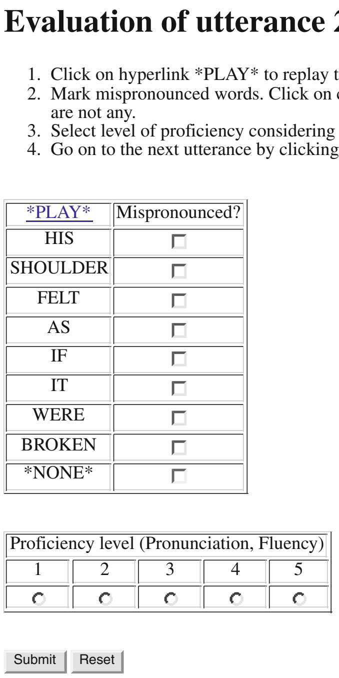
Figure D.3 from [Cincarek 04a] shows the web-browser based evaluation screen the human raters used for accessing and rating the speech. The data was transmitted via the HTTP protocol to a script which stored the ratings in a file.

Table D.1 Information about the evaluators

Evaluator ID	Working places	Teaching experience	Background in phonetics
2	Private school (Japan)	9 months	No
3	Private school (Japan)	6 months	Yes
5	Private school/companies (Japan)	5 years	Yes
6	Private school/companies/privately	4–5 years	No
7	Privately (Japan)	2–3 years	Some
8	Privately (Japan)	1–2 years	Yes
9	University (Japan)	5 years	Yes
10	Public school (Japan)	3 years	Some
11	Private school (Japan, Canada)	2–3 years	Yes
12	Private school (Japan)	17 months	No
13	Private school (Japan)	18 months	No
15	Privately (Japan)	3-4 years	No
16	Private schools/companies (Taiwan, Hong Kong, Japan, Canada)	8 years	No
17	Company (Japan)	6 years	Some
20	Various places (Europe,Asia) Creation of teaching material (e.g. audio CDs for TOIEC test)	9 years	Yes

Table D.2 Which rater evaluated which non-native speaker?

Evaluator IDs	Non-native speaker IDs
8 12 16 20	M082 M027 M077 F018 M044 M088 M010 M080 M072 M064 M030 M049 M025 M033 M093 F012 M023 F014 M014 M055 M061 M028 M056 M076
5 9 13 17	M029 M021 M038 M057 F021 M047 M066 F023 M011 M054 F009 M075 M031 M034 M067 M051 M015 M032 M062 M006 F008 M065 M090 M086
2 6 10	F010 M071 M052 M045 M024 M081 F025 M092 M016 M078 M073 M040 F013 M046 M053 F022 M084 M091 M039 F019 M013 M036 M050 M037
3 7 11 15	M074 M043 M026 M070 M060 F024 F026 M022 M041 F011 M087 M058 M042 M069 M012 F020 M085 M068 M089 M063 M059 M083 M001 M035



Evaluation of utterance 2 of 1152

1. Click on hyperlink *PLAY* to replay the utterance.
2. Mark mispronounced words. Click on checkbox for *NONE* if there are not any.
3. Select level of proficiency considering pronunciation and fluency.
4. Go on to the next utterance by clicking on the submit button.

PLAY	Mispronounced?
HIS	☐
SHOULDER	☐
FELT	☐
AS	☐
IF	☐
IT	☐
WERE	☐
BROKEN	☐
NONE	☐

Proficiency level (Pronunciation, Fluency)				
1	2	3	4	5
○	○	○	○	○

Glossary

AM	Acoustic model
ASCII	American standard code for information interchange
ASR	Automatic speech recognition
ATR	Advanced Telecommunication Research Laboratories
C-STAR	Consortium for speech translation advancedresearch
CALL	Computer assisted language learning
CART	Classification and regression tree
CMS	Cepstral mean substraction
DAT	Digital audio tape
ELDA	Evaluations and Language Resources DistributionAgency
ELRA	European Language Resource Agency
EM	Expectation maximization
ERJ	English read by Japanese
FSA	Finite state automaton
GMM	Gaussian mixture model
GUI	Graphical user interface
HMM	Hidden Markov model
IPA	International phonetic alphabet
ISLE	Interactive spoken language education project
L1	Native language of a speaker

L2	An aquired language a person is speaking with a non-native accent
LM	Language model
M-ATC	Military Air Traffic Control
MAP	Maximum a-posteriori
MFCC	Mel frequency cepstral coefficients
MIT	Massachusetts Institute of Technology
MLLR	Maximum likelihood linear regression
MWC	Multilingual weighted codebooks
NATO	North Atlantic Treaty Organisation
NICT	National Institute of Information and Communications Technology
PBC	Phoneme best correctness
PC	Personal computer
PCA	Principal component analysis
PDA	Personal digital assistant
PNA	Phoneme net accuracy
ROS	Rate of speech
SAMPA	Speech assessment methods phonetic alphabet
SDS	Speech dialog system
SMILE	Speech and multimodal interface for multilingual exchange
SPINE	Speech in Noisy Environments
TED	Translanguage English database
TIMIT	A database created by Texas Instruments and MIT
TTS	Text to speech
VQ	Word error rate
WPP	Word posterior probability
WRR	Weighted recognition rate
WSJ	Wall Street Journal

References

[Akahane-Yamada 04] Reiko Akahane-Yamada, Hiroaki Kato, Takahiro Adachi, Hideyuki Watanabe, Ryo Komaki, Rieko Kubo, Tomoko Takada, Yuko Ikuma, Hiroaki Tagawa, Keiichi Tajima and Hideki Kawahara. ATR CALL: A speech perception/production training system utilizing speech technology. In Proc. ICA, 2004.

[Amdal 00a] Ingunn Amdal, Filipp Korkmazskiy and ArunC Surendran. Joint Pronunciation Modeling of Non-Native Speakers Using Data-Driven Methods. Proc. ICSLP, 622–625, 2000.

[Amdal 00b] Ingunn Amdal, Filipp Korkmazskiy and Arun C. Surendran. Joint pronunciation modelling of non-native speakers using data-driven methods. In Proc. ICSLP, Beijing, China, 622–625, 2000.

[Arslan 97] Levent M. Arslan and John H.L. Hansen. Frequency Characteristics of Foreign Accented Speech. In Proc.ICASSP, Munich, Germany, 1123–1126, 1997.

[Association 99] International Phonetic Association. Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet. (Cambridge University Press, Cambridge, 1999).

[Baum 66] L.E. Baum and T.Petrie. Statistical Inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics, vol. 37, pages 1559–1563, 1966.

[Benarousse 99] Laurent Benarousse et al. The NATO Native and Non-Native (N4) Speech Corpus. In Proc. of the MIST workshop (ESCA-NATO), Leusden, September 1999.

[Bernstein 90] Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev and Mitchel Weintraub. Automatic evaluation and training in English pronunciation. In Proc. ICSLP, 1990.

[Billa 97] Jayadev Billa, Kristine Ma, John McDonough, George Zavaliagkos and David Miller. Multilingual Speech Recognition: The 1996 Byblos CALLHOME System. In Proc. EuroSpeech, 363–366, 1997.

[Binder 01] Norbert Binder, Konstantin Markov, Rainer Gruhn and Satoshi Nakamura. Speech Non-Speech Separation with GMMs. Proc. Acoust. Soc. Jap., vol. 1, 141–142, Fall 2001.

[Binder 02] Norbert Binder, Rainer Gruhn and Satoshi Nakamura. Recognition of Non-Native Speech Using Dynamic Phoneme Lattice Processing. Proc. Acoust. Soc. Jap., p. 203f, 2002. spring meeting.

[Blackburn 95] Simon Blackburn and Steven J. Young. Towards improved speech recognition using a speech production model. In Proc. EuroSpeech, volume 2, pages 1623–1626, 1995.

[Bronstein 07] Ilja N. Bronstein and Konstantin A. Semendjajew. Handbook of mathematics. Springer, 2007.

[Byrne 98] William Byrne, Eva Knodt, Sanjeev Khudanpur and Jared Bernstein. Is automatic speech recognition ready for non-native speech? A data-collection effort and initial experiments in modeling conversational Hispanic English. In STiLL, Marholmen, Sweden, 37–40, 1998.

[C-Star 99] C-Star. Consortium for Speech Translation Advanced Research (C-Star): C-Star Experiment. http://www.c-star.org/ July 1999.

[Campbell 00] Nick Campbell. Databases of Expressive Speech. In Proc ISCA (International Speech Communication and Association) ITRW in Speech and Emotion, 34–38, 2000.

[Campbell 02] Nick Campbell. Recording techniques capturing natural every-day speech. In Proc LREC, 2002.

[Cincarek 04a] Tobias Cincarek. Pronunciation scoring for non-native speech. Master's thesis, Univ. Erlangen-Nuernberg, 2004.

[Cincarek 04b] Tobias Cincarek, Rainer Gruhn and Satoshi Nakamura. Pronunciation scoring and extraction of mispronounced words for non-native speech. Proc. Acoust. Soc. Jap., 165–166, Fall 2004. (in Japanese).

[Cincarek 09] Tobias Cincarek, Rainer Gruhn, Christian Hacker, Elmar Noeth and Satoshi Nakamura. Automatic Pronunciation Scoring of Words and Sentences Independent from the Non-Native's First Language. Computer Speech and Language. 23 (2009).

[Cox 02] Stephen Cox and Srinandan Dasmahapatra. High-level approaches to confidence estimation in speech recognition. IEEE Transactions on Speech and Audio Processing, vol. 10, no. 7, pages 460–471, 2002.

[Cremelie 97] Nick Cremelie and Jean-Pierre Martens. Rule-based generation of word pronunciation networks. In Proc. EuroSpeech, 2459–2462, 1997.

[Cucchiarini 93] Catia Cucchiarini. Phonetic Transcription: a methodological and empirical study. PhD thesis, University of Nijmegen, 1993.

[FAA 07] FAA. Federal Aviation Administration Controller Pilot Datalink Communications (CPDLC). http://tf.tc.faa.gov/capabilities/cpdlc.htm, 2007.

[Fischer 03] Volker Fischer, Eric Janke and Siegfried Kunzmann. Recent progress in the decoding of non-native speech with multilingual acoustic models. In Proc.Eurospeech, 3105–3108, 2003.

[Fisher 87] William Fisher, Victor Zue, Jared Bernstein and David Pallet. An Acoustic–Phonetic Data Base. J. Acoust. Soc. Am. 81 (1987).

[Fisher 96] William Fisher. NIST tsylb2 syllabification software. ftp://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z, 1996.

[Fitt 95] Susan Fitt. The pronunciation of unfamiliar native and non-native town names. In Proc. Eurospeech, 2227–2230, 1995.

[Franco 00] Horacia Franco, Leonardo Neumeyer, Vassilios Digalakis and Orith Ronen. Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, vol. 30, pages 121–130, 2000.

[Fujisaki 96] Hiroya Fujisaki. Prosody, Models, and Spontaneous Speech. In Norio Higuchi Yoshinori Sagisaka Nick Campbell, editor, Computing Prosody, chapter 3, 27–42. Springer, 1996.

[Fukada 97] F. Fukada and Y. Sagisaka. Automatic Generation of a Pronunciation Dictionary Based on Pronunciation Networks. IEICE TRANSACTIONS, vol. J80-D-II, no.10, 2626–2635 October 1997.

[Goronzy 01] Silke Goronzy, Marina Sahakyan and Wolfgang Wokurek. Is Non-Native Pronunciation Modeling Necessary? Proc. EuroSpeech, 2001.

[Goronzy 04] Silke Goronzy, Stefan Rapp and Ralf Kompe. Generating non-native pronunciation variations for lexicon adaptation. Speech Communication, vol. 42, no. 1, pages 109–123, 2004.

[Gruhn 98] Rainer Gruhn. Adaption des VERBMOBIL-Spracherkenners und des Prosodiemoduls an die japanische Sprache. Master's thesis, Friedrich-Alexander-University Erlangen-Nuremberg, 1998.

[Gruhn 99] Rainer Gruhn, Harald Singer and Yoshinori Sagisaka. Scalar Quantization of Cepstral Parameters for Low Bandwidth Client-Server Speech Recognition Systems. In Proc. Acoust. Soc. Jap., 129–130 November 1999.

[Gruhn 00a] Rainer Gruhn, Satoshi Nakamura and Yoshinori Sagisaka. Towards a Cellular Phone Based Speech-To-Speech Translation Service. In Proc. MSC workshop on Multilingual Speech Communication, Kyoto, 2000.

[Gruhn 00b] Rainer Gruhn, Harald Singer, Hajime Tsukada, Atsushi Nakamura, Masaki Naito, Atsushi Nishino, Yoshinori Sagisaka and Satoshi Nakamura. Cellular Phone Based Speech-To-Speech Translation System ATR-MATRIX. In Proc. ICSLP, 448–451, 2000.

[Gruhn 01a] Rainer Gruhn and Satoshi Nakamura. Multilingual Speech Recognition with the CALLHOME Corpus. In Proc. Acoust. Soc. Jap., vol. 1, 153–154, Fall 2001.

[Gruhn 01b] Rainer Gruhn, Koji Takashima, Takeshi Matsuda, Atsushi Nishino and Satoshi Nakamura. CORBA-based Speech-to-Speech Translation System. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 355–358 December 2001.

[Gruhn 01c] Rainer Gruhn, Koji Takashima, Takeshi Matsuda, Atsushi Nishino and Satoshi Nakamura. A CORBA-based Speech-to-Speech Translation System. In Proc. Acoust. Soc. Jap., 153–154, Fall 2001.

[Gruhn 02] Rainer Gruhn, Konstantin Markov and Satoshi Nakamura. Probability Sustaining Phoneme Substitution for Non-Native Speech Recognition. Proc. Acoust. Soc. Jap., 195–196, Fall 2002.

[Gruhn 04a] Rainer Gruhn, Tobias Cincarek and Satoshi Nakamura. A Multi-Accent Non-Native English Database. In Proceedings of the Acoustic Society of Japan, 163–164, 2004.

[Gruhn 04b] Rainer Gruhn, Konstantin Markov and Satoshi Nakamura. Discrete HMMs for statistical pronunciation modeling. SLP 52 / HI 109, 123–128, 2004.

[Gruhn 04c] Rainer Gruhn, Konstantin Markov and Satoshi Nakamura. A statistical lexicon for non-native speech recognition. In Proc. of Interspeech, (Jeju Island, Korea, 2004) 1497–1500.

[Gruhn 04d] Rainer Gruhn and Satoshi Nakamura. A statistical lexicon based on HMMs. Proc. IPSJ, 2, p. 37f, 2004.

[Hacker 07] Christian Hacker, Tobias Cincarek, Andreas Maier, Andre Hessler and Elmar Noeth. Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children. In Proc. ICASSP, (Honolulu, Hawai, 2007) 197–200.

[Hanson 97] Helen M. Hanson. Glottal characteristics of female speakers: Acoustic correlates. In Acoustical Society of America, 1997.

[Hirsch 00] Hans-Guenther Hirsch and David Pearce. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In ISCA ITRW ASR 2000, September 2000.

[Huang 01] Xuedong Huang, Alejandro Acero and Hsiao-Wuen Hon. Spoken language processing: A guide to theory, algorithm, and system development. (Prentice Hall, 2001).

[Institute 07] TNO Human Factors ∼ Research Institute. MIST Multi-lingual Interoperability in Speech Technology database. Technical report, ELRA, Paris, France, 2007. ELRA Catalog Reference S0238.

[Jelinek 97] Frederick Jelinek. Statistical methods for speech recognition. (MIT Press, Cambridge, 1997).

[Junqua 94] Jean-Claude Junqua, Brian Mak and Ben Reaves. A Robust Algorithm for Word Boundary Detection in the Presence of Noise. Proc. IEEE, vol. 2, no. 3, 406–412, 1994.

[Jurafsky 94] Daniel Jurafsky, Chuck Wooters, Gary Tajchman, Jonathan Segal, Andreas Stolcke, Eric Fosler and Nelson Morgan. The Berkeley Restaurant Project. In Proc. ICSLP, 1994.

[Kato 04] Hiroaki Kato, Keiichi Tajima, Amanda Rothwell, Reiko Akahane-Yamada and Kevin Munhall. Perception of phonemic length contrasts in Japanese with or without a carrier sentence by native and non–native listeners. In Proc. ICA, 2004.

[Kawaii 99] Goh Kawaii. Spoken language processing applied to nonnative language pronunciation learning. PhD thesis, University of Tokyo, 1999.

[Kiessling 92] Andreas Kiessling, Ralf Kompe, Heinrich Niemann and Elmar Noeth. DP-Based Determination of F0 Contours from Speech Signals. In Proc. ICASSP, 1992.

[Kim 07] Mina Kim, Yoo Rhee Oh and Hong Kook Kim. Non-Native Pronunciation Variation Modeling Using an Indirect Data Driven Method. In Proc. ASRU, 2007.

[Köhler 96] Joachim Köhler. Multilingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds. Proc. ICSLP, vol. 4, pages 2195–2198, 1996.

[Lamel 94] Lori F. Lamel, Florian Schiel, Adrian Fourcin, Joseph Mariani and Hans G. Tillmann. The Translanguage English Database TED. In Proc. ICSLP, Yokohama, Japan, September 1994.

[Lander 07] Terri Lander. CSLU: Foreign Accented English Release 1.2. Technical report, LDC, Philadelphia, Pennsylvania, 2007.

[LaRocca 02] Stephen A. LaRocca and Rajaa Chouairi. West Point Arabic Speech Corpus. Technical report, LDC, Philadelphia, Pennsylvania, 2002.

[LaRocca 03] Stephen A. LaRocca and Christine Tomei. West Point Russian Speech Corpus. Technical report, LDC, Philadelphia, Pennsylvania, 2003.

[Livescu 99] Karen Livescu. Analysis and modeling of non-native speech for automatic speech recognition. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[Markov 03] Konstantin Markov, Tomoko Matsui, Rainer Gruhn, Jinsong Zhang and Satoshi Nakamura. Noise and Channel Distortion Robust ASR System for DARPA SPINE2 Task. IEICE Transactions on Information and Systems, vol. E86-D No.3, 497–504, March 2003.

[Masataki 97] Hirokazu Masataki, Kazoku Takahashi, Hajime Tsukada, Kouichi Tanigaki, Hirofumi Yamamoto, Atsushi Nishino and Atsushi Nakamura. Baseline Language Model for Continuous Speech Recognition. Technical report TR-IT-0235, ATR, 1997. (in Japanese).

[Mengistu 08] Kinfe Tadesse Mengistu and Andreas Wendemuth. Accent and Channel Adaptation for Use in a Telephone-based Spoken Dialog System. In Proc. TSD, 2008.

[Menzel 00] Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, R. Morton and C. Souter. The ISLE corpus of non-native spoken English. In Proc. LREC, pages 957–963, Athens, Greece, 2000.

[Minematsu 02] Nobuaki Minematsu, Yoshihiro Tomiyama, Kei Yoshimoto, Katsumasa Shimizu Seiichi Nakagawa, Masatake Dantsuji and Shozo Makino. English Speech Database Read by Japanese Learners for CALL System Development. In Proc.LREC, Las Palmas, Spain, 2002.

[Minematsu 03] Nobuaki Minematsu, Gakuro Kurata and Keikichi Hirose. Improvement of non-native speech recognition by effectively modelling frequently observed pronunciation habits. Proc. EuroSpeech, 2597–2600, 2003.

[Minematsu 04] Nobuaki Minematsu, Yoshihiro Tomiyama, Kei Yoshimoto, Katsumasa Shimizu ∼ Seiichi Nakagawa, Masatake Dantsuji and Shozo Makino. Development of English Speech Database read by Japanese to Support CALL Research. In Proc. ICA, vol. 1 (Kyoto, Japan, 2004) 577–560.

[Minker 98a] Wolfgang Minker. Evaluation Methologies for Interactive Speech Systems. In Proc. LREC, 1998.

[Minker 98b] Wolfgang Minker. Stochastic versus Rule-based Speech Understanding for Information Retrieval. Speech Communication, vol. 25(4), 1998.

[Minker 02a] Wolfgang Minker. Overview on Recent Activities in Speech Understanding and Dialogue Systems Evaluation. In Proc. ICSLP, Denver, 2002.

[Minker 02b] Wolfgang Minker, Udo Haiber, Paul Heisterkamp and Sven Scheible. Intelligent Dialogue Strategy for Accessing Infotainment Applications in Mobile Environments. In Proc. ISCA Workshop on Multi-modal Dialogue in Mobile Environments, Kloster Irsee, 2002.

[Morgan 06] John Morgan. West Point Heroico Spanish Speech. Technical report, LDC, Philadelphia, Pennsylvania, 2006.

[Mote 04] Nicolaus Mote, Lewis Johnson, Abhinav Sethy, Jorge Silva and Shrikanth Narayanan. Tactical Language Detection and Modeling of Learner Speech Errors: The case of Arabic tactical language training for American English speakers. In Proc. of InSTIL, June 2004.

[Munich 98] University Munich. Bavarian Archive for Speech Signals Strange Corpus, 1998. http://www.phonetik.uni-muenchen.de/Bas/.

[Munich 00] University Munich. The Verbmobil Project, 2000. http://www.phonetik.uni-uenchen.de/Forschung/Verbmobil/VerbOverview.html.

[Murai 00] Kazumasa Murai, Rainer Gruhn and Satoshi Nakamura. Speech Start/End Point Detection Using Mouth Image. Proc. IPSJ, vol. 2, 169–170, Fall (2000).

[Nakamura 96] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura and Y. Sagisaka. Japanese Speech Databases for Robust Speech Recognition. Proc. ICSLP, 2199–2202, 1996.

[Neri 04] Ambra Neri, Catia Cucchiarini and Helmer Strik. Feedback in Computer Assisted Pronunciation Training: Technology Push or Demand Pull? In Proc. ICSLP, 1209–1212, 2004.

[Neumeyer 00] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis and Mitchel Weintraub. Automatic Scoring of pronunciation quality. Speech Communication, vol. 30, 83–93 (2000).

[Nishina 04] Kikuko Nishina, Yumiko Yoshimura, Izumi Saita, Yoko Takai, Kikuo Maekawa, Nobuaki Minematsu, Seiichi Nakagawa, Shozo Makino and Masatake Dantsuji. Development of Japanese Speech Database Read by Non-native Speakers for Constructing CALL System. In Proc. ICA, pages 561–564, Kyoto, Japan, 2004.

[Nishiura 01a] Takanobu Nishiura, Rainer Gruhn and Satoshi Nakamura. Automatic Steering of Microphone Array and Video Camera Toward Multi-Lingual Tele-Conference Through Speech-to-Speech Translation. In Proc. IEEE International Conference on Multimedia and Expo (ICME2001), 569–572 August 2001.

[Nishiura 01b] Takanobu Nishiura, Rainer Gruhn and Satoshi Nakamura. Collaborative Steering of Microphone Array and Video Camera Toward Multi-Lingual Tele-Conference Through Speech-to-Speech Translation. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 2001.

[Nishiura 02a] Takanobu Nishiura, Rainer Gruhn and Satoshi Nakamura. Automatic Steering of Microphone Array and Video Camera Toward Multi-Lingual Tele-Conference Through Speech-to-Speech Translation. The Journal of the Information Processing Society of Japan, 3617–3620 December 2002.

[Nishiura 02b] Takanobu Nishiura, Rainer Gruhn and Satoshi Nakamura. A Prototype System Design of Distant Talking Speech Translation with a Microphone Array and Video Camera. In Proc. Acoust. Soc. Jap., vol. 1, 229–230, Spring 2002.

[Noeth 90] Elmar Noeth. Prosodische Information in der automatischen Sprachverarbeitung, Berechnung und Anwendung. Niemayer, Tübingen, 1990.

[OMG 90] OMG. Common Object Request Broker Architecture (CORBA). http://www.corba.org/, 1990.

[Onomastica 95] Consortium Onomastica. The ONOMASTICA interlanguage pronunciation lexicon. In Proc. Eurospeech, pages 829–832, Madrid, Spain, 1995.

[Paliwal 90] K.K. Paliwal. Lexicon-Building Methods for an Acoustic Sub-Word based Speech Recognizer. In Proc. ICASSP, 729–732, 1990.

[Paul 92] Dougla B. Paul and Janet∼M. Baker. The Design for the Wall Street Journal based CSR Corpus. In Proc. DARPA Workshop, 357–362, Pacific Grove, CA, 1992.

[Pigeon 07] Stephane Pigeon, Wade Shen and David van Leeuwen. Design and characterization of the Non-native Military Air Traffic Communications database. In Proc.ICSLP, Antwerp, Belgium, 2007.

[Raab 07] Martin Raab, Rainer Gruhn and Elmar Noeth. Non-native speech databases. In Proc. ASRU, 413–418, 2007.

[Raab 08a] Martin Raab, Rainer Gruhn and Elmar Noeth. Codebook Design for Speech Guided Car Infotainment Systems. In Proc. PIT, 44–51, Kloster Irsee, 2008.

[Raab 08b] Martin Raab, Rainer Gruhn and Elmar Noeth. Multilingual Weighted Codebooks. In Proc. ICASSP, 4257–4260, 2008.

[Raab 08c] Martin Raab, Rainer Gruhn and Elmar Noeth. Multilingual Weighted Codebooks for Non-native Speech Recognition. In Proc. TSD, 485–492, 2008.

[Raab 11] Martin Raab, Rainer Gruhn and Elmar Noeth. A scalable architecture for multilingual speech recognition on embedded devices. Speech Communication, vol. 53, no. 1, pages 62–74, 2011.

[Rabiner 75] Lawrence Richard Rabiner and MMarvin Robert Sambur. An algorithm for determining the endpoints of isolated utterances. The Bell System Technical Journal, vol. 54, 297–315, February (1975).

[Reaves 93] Ben Reaves. Parameters for Noise Robust Speech Detection. In Proc. Acoust. Soc. Jap., 145–146, Fall (1993).

[Rhee 04] S-C. Rhee, S-H. Lee, S-K. Kang and Y-J. Lee. Design and Construction of Korean-Spoken English Corpus (K-SEC). In Proc. ICSLP, 2004.

[Riley 99] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters and G. Zavaliagkos. Stochastic pronunciation modeling from hand-labelled phonetic corpora. Speech Communication, vol. 29, pages 209–224, 1999.

[Sakti 07] Sakriani Sakti, Konstantin Markov and Satoshi Nakamura. An HMM Acoustic Model Incorporating Various Additional Knowledge Sources. In Proc. EuroSpeech, 2117–2120, 2007.

[Schaden 02] Stefan Schaden. A Database for the Analysis of Cross-Lingual Pronunciation Variants of European City Names. In Proc.LREC, Las Palmas de Gran Canaria, Spain, 2002.

[Schaden 06] Stefan Schaden. Regelbasierte Modellierung fremdsprachlich akzentbehafteter Aussprachevarianten. PhD thesis, University Duisburg-Essen, 2006.

[Schukat-Talamazzini 95] Ernst-Günther Schukat-Talamazzini. Automatische Spracherkennung. Vieweg, Wiesbaden, 1995.

[Schultz 00a] Tanja Schultz. Multilinguale Spracherkennung. PhD thesis, Universität Karlsruhe, 2000.

[Schultz 00b] Tanja Schultz and Alex Waibel. Experiments towards a multi-language LVCSR interface. In Proc. ICSLP, 129–132, Bejing, China, 2000.

[Schultz 01] Tanja Schultz and Alex Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Communication, vol. 35, pages 31–51, 2001.

[Segura 07] Jose C. Segura et al. The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication, 2007. http://www.hiwire.org/.

[Seneff 98] Stephanie Seneff, Ed Hurley, Raymond Lau, Christine Pao, Philipp Schmid and Victor Zue. Galaxy-II: A Reference Architecture for Conversational System Development. In Proc. ICSLP, Sydney, 1998.

[Shoup 80] J.E. Shoup. Phonological aspects of speech recognition. In Trends in Speech Recognition, 125–138, 1980.

[Singer 99a] Harald Singer, Rainer Gruhn and Yoshinori Sagisaka. Speech Translation Anywhere: Client-Server Based ATR-MATRIX. In Proc. Acoust. Soc. Jap., 165–166 November 1999.

[Singer 99b] Harald Singer and Hirufumi Yamamoto. Parallel Japanese/English Speech Recognition in ATR-SPREC. In Proc. Acoust. Soc. Jap., 167–168, 1999.

[Sloboda 96] Tilo Sloboda and Alex Waibel. Dictionary learning for Spontaneous Speech Recognition. In Proc. ICSLP, 2328–2331, 1996.

[Steidl 04] Stefan Steidl, Georg Stemmer, Christian Hacker and Elmar Noeth. Adaption in the Pronunciation Space for Nonnative Speech Recognition. In D. Kim S. and Youn, editor, Proc. ICSLP, pages 318–321, Jeju Island, Korea, 2004.

[Stemmer 01] Georg Stemmer, Elmar Noeth and Heinrich Niemann. Acoustic Modeling of Foreign Words in a German Speech Recognition System. In P.Dalsgaard, B. Lindberg and H. Benner, editors, Proc. Eurospeech, volume 4, 2745–2748, 2001.

[Strik 99] Helmer Strik and Catia Cucchiarini. Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication, vol. 29 225–246 (1999).

[Sugaya 00] F. Sugaya, T. Takezawa, A. Yokoo, Y. Sagisaka and S. Yamamoto. Evaluation of the ATR-MATRIX speech translation system with a pair comparison methodbetween the system and humans. In Proc. ICSLP, Oct 2000.

[Sun 06] Ying Sun, Daniel Willett, Raymond Brueckner, Rainer Gruhn and Dirk Buehler. Experiments on Chinese Speech Recognition with Tonal Models and Pitch Estimation Using the Mandarin Speecon Data. In Proc. ICSLP, pages 1245–1248, Pittsburgh, 2006.

[Svojanovsky 04] Wenzel Svojanovsky, Rainer Gruhn and Satoshi Nakamura. Classification of nonverbal utterances in Japanese spontaneous speech. Proc. IPSJ, vol. 2, p. 277f, 2004.

[Tan 06] Tien-Ping Tan and Laurent Besacier. A French Non-Native Corpus for Automatic Speech Recognition. In Proc. LREC, Genoa, Italy, 2006.

[Tanigaki 00] Koichi Tanigaki, Hirofumi Yamamoto and Yoshinori Sagisaka. A Hierarchical Language Model Incorporating Class-dependent Word Models for OOV Word Recognition. In Proc. ICSLP, 2000.

[Teixeira 97] Carlos Teixeira, Isabel Trancoso and Antonio Serralheiro. Recognition of Non-Native Accents. In Proc. Eurospeech, pages 2375–2378, Rhodes, Greece, 1997.

[Teixeira 00] Carlos Teixeira, Horacio Franco, Elizabeth Shriberg, Kristin Precoda and Kemal Sönmez. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In Proc. ICSLP, volume 3, pages 187–190, 2000.

[Tomokiyo 00] Laura Mayfield Tomokiyo. Lexical And Acoustic Modeling of Non Native Speech in LVCSR. Proc. ICSLP, pages IV:346–349, 2000.

[Tomokiyo 01] Laura Tomokiyo. Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in Speech Recognition. PhD thesis, Carnegie Mellon University, Pennsylvania, 2001.

[Trancoso 99] Isabel Trancoso, Ceu Viana, Isabel Mascarenhas and Carlos Teixeira. On Deriving Rules for Nativised Pronunciation in Navigation Queries. In Proc. Eurospeech, 1999.

[Übler 98] Ulla Übler, Michael Schüßler and Heinrich Niemann. Bilingual And Dialectal Adaptation And Retraining. Proc. ICSLP, 1998.

[Übler 01] Ulla Übler. Multilingual speech recognition in seven languages. Speech Communication, no. 35, 53–69, 2001.

[Umeno 03] Junji Umeno. Analysis of Para-linguistic and Non-linguistic Information Based on Acoustic Parameters of Aizuchi – ‘un‘ in the Conversation. Master’s thesis, Nara Institute of Science and Technology, Nara, 2003.

[van Compernolle 01] Dirk van Compernolle. Recognition of goats, wolves, sheep and … non-natives. Speech Communication, vol. 35, 71–79 (2001).

[van den Heuvel 06] Henk van den Heuvel, Khalid Choukri, Christian Gollan, Asuncion Moreno and Djamal Mostefa. TC-STAR: New language resources for ASR and SLT purposes. In Proc. LREC, 2570–2573, Genoa, 2006.

[Vasquez 08] Daniel Vasquez, Rainer Gruhn, Raymond Brueckner and Wolfgang Minker. Comparing Linear Feature Space Transformations for Correlated Features. In Proc. PIT, 176–187, Kloster Irsee, 2008.

[Viterbi 79] Andrew Viterbi. Principles of digital communication and coding. McGraw-Hill College, 1979.

[Waibel 88] Alex Waibel. Prosody and Speech Recognition. Pitman, 1988.

[Wang 03] Zhirong Wang, Tanja Schultz and Alex Waibel. Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech. In Proc. ICASSP, 2003.

[Wells 95] J.C. Wells. Computer-Coding the IPA: a porposed extension of SAMPA. Technical report, University College London, April 1995.

[Wessel 01] Frank Wessel, Ralf Schlüter, Klaus Macherey and Hermann Ney. Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing, vol. 9, no.3, pages 288–298, 2001.

[Wester 00] Mirjam Wester and Eric Fosler-Lussier. A comparison of data-derived and knowledge-based modeling of pronunciation variation. In Proc. ICSLP, 270–273, 2000.

[Wikipedia 07] Wikipedia. http://en.wikipedia.org/wiki/Non-native_speech_databases, 2007.

[Witt 99a] Silke Witt. Use of Speech Recognition in Computer-Assisted Language Learning. PhD thesis, Cambridge University Engineering Department, UK, 1999.

[Witt 99b] Silke Witt and Steve Young. Off-Line Acoustic Modeling of Non-Native Accents. Proc. EuroSpeech, 1367–1370, 1999.

[Woodland 93] Phil Woodland and Steve Young. The HTK Tied-State Continuous Speech Recognizer. In Proc. EuroSpeech, 2207–2210, 1993.

[Yang 00] Qian Yang and Jean-Pierre Martens. Data-driven lexical modeling of pronunciation variations for ASR. In Proc. ICSLP, 417–420, 2000.

[Yang 02] Qian Yang, Jean-Piere Martens, Pieter-Jan Ghesquiere and Dirk van Compernolle. Pronunciation Variation Modeling for ASR: Large Improvements are Possible but Small Ones are Likely to Achieve. In Proc. PMLA, 123–128, 2002.

[Ye 05] Hui Ye and Steve Young. Improving the Speech Recognition Performance of Beginners in Spoken Conversational Interaction for Language Learning. In Proc. Interspeech, Lisbon, Portugal, 2005.

[Yun 99] Seong-Jin Yun and Yung-Hwan Oh. Stochastic Lexicon Modeling for Speech Recognition. IEEE Signal Processing Letters, vol. 6, 28–30 (1999).

[Zhang 02] Jinsong Zhang, Konstantin Markov, Tomoko Matsui, Rainer Gruhn and Satoshi Nakamura. Developing Robust Baseline Acoustic Models for Noisy Speech Recognition in SPINE2 Project. In Proc. Acoust. Soc. Jap., volume 1, 65–66 Spring 2002.