Normalization in the Acoustic Feature Space for Improved Speech Recognition

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Diplom–Informatiker Sirko Molau

aus

Berlin

Berichter: Universitätsprofessor Dr.–Ing. Hermann Ney
Universitätsprofessor Dr. phil. nat. Harald Höge

Tag der mündlichen Prüfung: 14. Februar 2003

To Eva and Hans-Joachim Molau.

Acknowledgements

At this point, I would like to express my thanks to the people who supported and accompanied me during the progress of this work.

I would like to thank my supervisor, Prof. Dr.-Ing. Hermann Ney, head of the Chair of Computer Science VI at the Technical University of Aachen, for giving me a deep insight into pattern recognition in general, and for introducing me to the field of automatic speech recognition in particular. He gave me the opportunity to pursue my ideas, he followed my work with continuous interest, and he supported me with numerous ideas and discussions.

I am also grateful to my second supervisor, Prof. Dr. phil. nat. Harald Höge from Siemens AG Munich, for his interest in this thesis and his valuable advice.

My experiments in vocal tract length normalization were based on the initial implementation by Lutz Welling, Nils Haberland and Stephan Kanthak. The research in histogram normalization and feature space rotation gained in particular from discussions with Florian Hilger and Daniel Keysers. Michael Pitz, Ralf Schlüter and Achim Sixtus especially supported my work all the time. I am very much indebted to all of them.

I would like to thanks my officemates Michael Motter and Oliver Bender for being great companions and good friends, as well as for their excellent support of our computing equipment.

I am indebted to all my colleagues at the Lehrstuhl für Informatik VI for their fruitful discussions, comments and hints, and for their patience when they shared their computing resources with me. It was a great pleasure to be a member of such an excellent team.

There were a number of frustrating days, when the experimental results were not as good as they were supposed to be, or when project work and teaching left too little time for research. Lunch time with my colleagues, however, was always a fine opportunity to provide some relaxation and "refresh the batteries" thanks to Achim, Andras, Frank, Michael, Nicola, Ralf, Sonja and Stephan.

Finally I would like to thank my family and in particular my girl friend Carina Theiler and our daughter Vanessa for their understanding, their support and the many months they had to live without me. I am indebted to my parents Eva and Hans-Joachim Molau for enabling and supporting me with my work and my interests. It goes without saying that without them, this thesis would not have been possible.

Abstract

In this work, normalization techniques in the acoustic feature space are studied which improve the robustness of automatic speech recognition systems.

It is shown that there is a fundamental mismatch between training and test data which causes degraded recognition performance. *Adaptation* and *normalization*, basic strategies to reduce the mismatch, are introduced and placed into the framework of statistical speech recognition. A classification scheme for different normalization techniques is introduced. Common normalization schemes proposed in the literature are motivated and discussed, and two promising techniques are implemented and studied in detail.

Vocal tract length normalization relies on frequency axis warping during signal analysis to reduce inter-speaker variability. The baseline procedure for training and test data normalization is introduced and optimized so that consistently large improvements in recognition performance are achieved under a variety of acoustic conditions. A technique for fast parameter estimation is developed that gives the same improvements as the baseline technique without an increase in computation time. It is shown that vocal tract length normalization can be applied successfully in online applications. A novel approach for integrated frequency axis warping is developed that merges successive signal analysis steps into a single one. It simplifies signal analysis and gives a better control over the amount of spectral smoothing.

The second set of techniques explored in detail are *histogram normalization* and *feature space rotation*. They aim at reducing the mismatch between training and test by matching the distributions of the training and test data. The effect of histogram normalization at different signal analysis stages, as well as training and test data normalization are investigated in detail. One of the basic assumptions of histogram normalization is relaxed by taking care of the variable silence fraction. Feature space rotation is introduced to account for undesired variations in the speech signal that are correlated in the feature space dimensions. The interaction of histogram normalization and feature space rotation is analyzed, and it is shown that both techniques significantly improve the recognition accuracy in scenarios with different degrees of mismatch.

Finally, it is demonstrated how the application of several normalization schemes in presence of large mismatch between training and test data can make the difference from essentially zero recognition accuracy to a high level of 90%.

Experimental results are reported for corpora with different acoustic conditions, vocabulary sizes, languages, and speaking styles: *North American Business News* is a large vocabulary task of English read speech, *VerbMobil II* is a German large vocabulary conversational speech task, *EuTrans II* is an Italian speech corpus of conversational speech over telephone, and *CarNavigation* a German isolated-word recognition task recorded partly in noisy car environments.

Zusammenfassung

In dieser Arbeit werden Normalisierungsverfahren im akustischen Merkmalsraum zur Erhöhung der Robustheit von automatischen Spracherkennungssystemen untersucht.

Es gibt eine grundsätzliche Diskrepanz zwischen den Trainings- und Testdaten, die zu einer Verschlechterung der Erkennungsleistung führt. *Adaption* und *Normalisierung*, zwei Prinzipien zur Verringerung des Unterschieds, werden in der Arbeit vorgestellt und in den Rahmen der statistischen Spracherkennung eingefügt. Es wird ein Klassifikationsschema für Normalisierungsverfahren entwickelt. Gängige Normalisierungsverfahren werden vorgestellt und erörtert und zwei besonders erfolgversprechende Verfahren im Rahmen der Arbeit umgesetzt und genauer analysiert.

Die *Vokaltraktlängennormierung* beruht auf der Verzerrung der Frequenzachse während der Signalanalyse mit dem Ziel, sprecherabhängige Variationen im Sprachsignal zu reduzieren. Das allgemeine Prinzip wird vorgestellt und das Standardverfahren so optimiert, daß konsistent hohe Verbesserungen der Erkennungsleistung in verschiedenen Umgebungen erreicht werden. Ein Verfahren zur schnellen Parameterschätzung liefert dieselben Verbesserungen ohne eine Zunahme an Rechenzeit, was den Einsatz der Normalisierung in Online-Erkennungssystemen ermöglicht. Schließlich wird ein neuer Ansatz zur integrierten Verzerrung der Frequenzachse vorgestellt, der mehrere Signalanalyseschritte zu einem vereint. Das vereinfacht die Signalanalyse und verbessert die Kontrolle über die spektrale Glättung.

Der zweite Satz von Verfahren, die im Detail untersucht werden, sind die *Histogrammnormalisierung* und die *Merkmalsraumrotation*. Sie zielen darauf ab, die Diskrepanz zwischen Trainings- und Testdaten durch eine Angleichung ihrer Verteilungen zu verringern. Der Effekt der Normalisierung auf verschiedenen Ebenen der Signalanalyse sowie auf Trainings- und Testdaten wird untersucht. Die Berücksichtigung des Anteils an Sprechpausen relaxiert eine der Grundannahmen der Histogrammnormalisierung. Ein Verfahren zur Merkmalsraumrotation beseitigt unerwünschte Variationen im Sprachsignal, die in den einzelnen Dimensionen des Merkmalsraumes korreliert sind. Die Interaktion von Histogrammnormalisierung und Rotation wird untersucht. Beide Verfahren erhöhen deutlich die Erkennungsleistung in Szenarien mit verschiedenen Graden an Diskrepanz zwischen Trainings- und Testdaten.

Schließlich wird demonstriert, daß die Anwendung mehrerer Normalisierungsverfahren im Fall von starker Diskrepanz zwischen Training und Test die Erkennungsleistung von Null auf ein hohes Niveau von 90% bringen kann.

Erkennungsergebnisse werden für Korpora mit verschiedenen akustischen Bedingungen, Vokabulargrößen, Sprachen und Sprechstilen angegeben: *North American Business News* ist ein Testkorpus mit großem Vokabular, der aus gelesenen englischen Texten besteht. *VerbMobil II* ist ein deutscher Spontansprachkorpus mit großem Vokabular, *EuTrans II* ist ein italienischer spontansprachlicher Telefonkorpus und *CarNavigation* ein verrauschter deutscher Einzelwortkorpus, der zum Teil in fahrenden Autos aufgenommen wurde.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Speech is the fundamental form of communication between humans, and it will play a central role in future man-machine interfaces. The aim of automatic speech recognition is to extract the sequence of spoken word from a recorded speech signal. It does not include the task of speech understanding, which can be seen as an even more elaborate problem.

Depending on the complexity of the recognition task, current automatic speech recognizers range from experimental systems in research laboratories to solutions that have reached the consumer market in a number of isolated applications. For speaker-dependent recognition of isolated digit strings in clean acoustic conditions, the word error rate can be as low as a fraction of a percent, whereas speaker-independent large vocabulary recognition of conversational speech in adverse acoustic conditions (e.g. the transcription of broadcast news) still yields word error rates in the order of several ten percent. Hence, today's best automatic speech recognizers are still far inferior to the performance of the human ear and brain.

Compared to the importance speech has in the communication among humans, the integration of voice in current man-machine-interfaces is still rudimentary. Examples for speech recognition systems that have reached the market are:

- voice control of technical devices (e.g. telephone, car electronics, ...)
- dictation systems (e.g. PC-based dictation systems)
- transcription systems (e.g. transcription of broadcast news, medical reports, ...)
- access to databases (e.g. information systems for time tables, telephone numbers, stock prices, voice-mail systems, ...)

Further research is required to design voice-based interfaces, and to improve the available speech recognition technology. A key issue is to increase the robustness of speech recognizers to variable environments and speakers. It will be of fundamental importance for the advance of speech technology.

1.1 Statistical Speech Recognition

Automatic speech recognition is nowadays solved in a statistical framework. Bayes' decision rule [Duda & Hart 73] states that the word sequence $W = w_1, \ldots, w_N$ should be chosen that maximizes the posterior probability of the observed sequence of acoustic vector $X = x_1, \ldots, x_N$:

$$W = \arg\max_{W'} p(W'|X) \tag{1.1}$$

Using Bayes' identity, the posterior probability can be transformed as:

$$p(W|X) = \frac{p(W) \cdot p(X|W)}{p(X)} \tag{1.2}$$

The a-priori probability $p(X)$ of the acoustic vector sequence can be omitted, since it is a constant factor and has no influence on the optimization problem. Hence, Bayes' decision rule can be rewritten as:

$$W = \arg\max_{W'} \{p(W') \cdot p(X|W')\} \tag{1.3}$$

From Eqn. 1.3 follows that there are two basic knowledge sources involved in automatic speech recognition: The acoustic model with the class-dependent probability distributions $p(X|W)$, and the language model that provides the a-priori probability $p(W)$ of the word sequence W. Both knowledge sources can be found in the system architecture depicted in Figure 1.1 [Ney 1990], which has become a de-facto standard for modern speech recognizer.

A speech recognition systems consists of four basic parts, which will be described in detail in the following sections:

- the *signal analysis* (Section 1.2) creates a sequence of acoustic vectors X from the speech waveform recorded by the microphone
- the *acoustic model* (Section 1.3) describes the probability to observe a sequence of acoustic vectors X given a (hypothesized) word sequence W. The acoustic model is typically made of two parts:
  - acoustic models for the smallest sub-word units that are used, i.e. typically phonemes
  - the pronunciation lexicon that describes how the acoustic model for words is composed from the sub-word units
- the *language model* (Section 1.4) covers syntax, semantics, and pragmatics of the language, and provides the a-priori probability of a (hypothesized) word sequence
- the *search* procedure (Section 1.5) finds the word sequence of maximum posterior probability according to Bayes' decision rule (Eqn. 1.3)

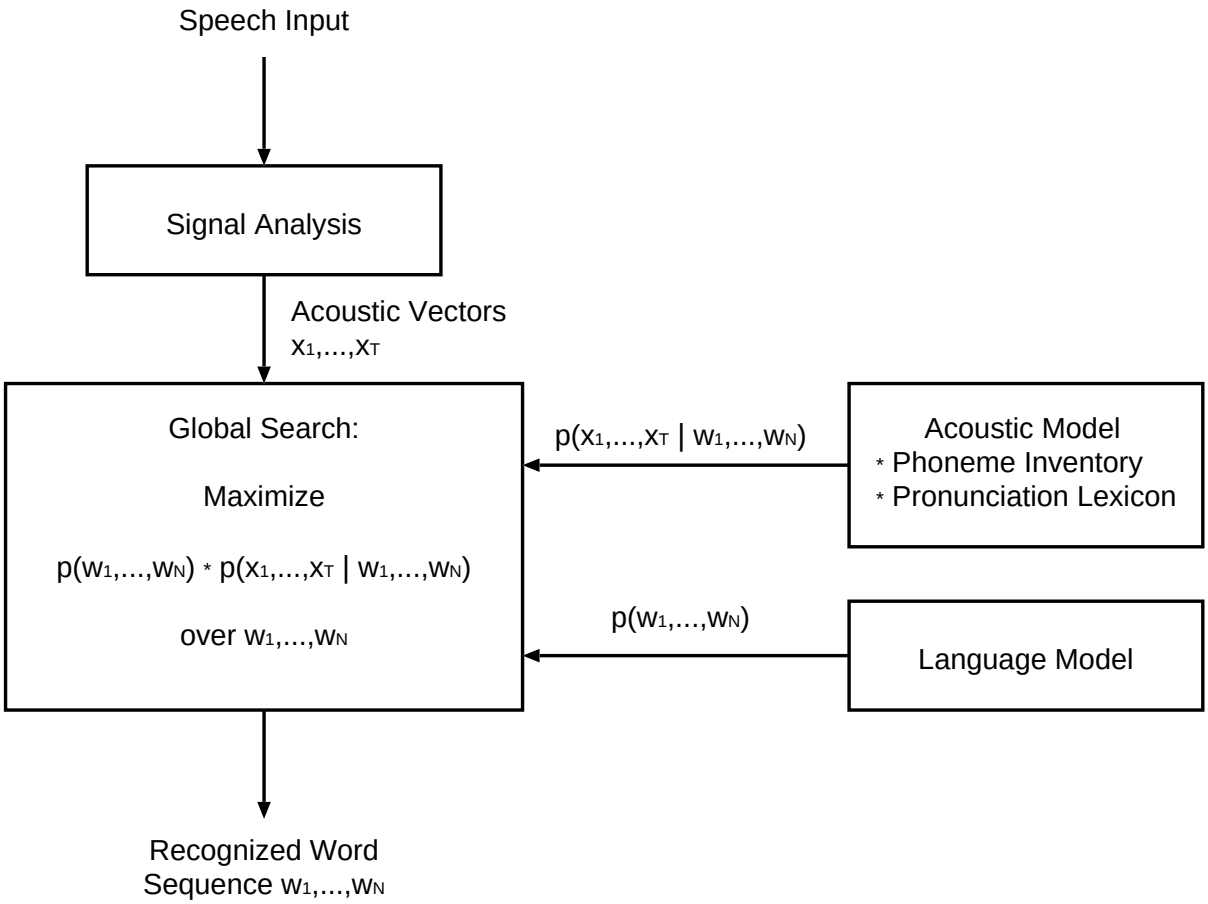

Figure 1.1: Principal architecture of an automatic speech recognition system.

1.2 Signal Analysis

The aim of signal analysis is to provide the speech recognizer with a stream of acoustic vectors. The vector sequence is a parameterization of the speech waveform observed at the microphone that should fulfill the following criteria:

- acoustic vectors should be characteristic for the sub-word units used in acoustic modeling, i.e. they should be similar for the same phoneme, but discriminative among different phonemes
- they should depend on the spoken word sequence only; ideally they are independent of the speaker, recording conditions, transmission channel, and other environmental effects
- the acoustic vectors should be of low dimensionality to allow robust parameter estimation for the acoustic model, i.e. all information necessary for the recognition process should be captured in only a few coefficients

Especially the second condition is difficult to realize, which is why special normalization and adaptation methods were developed to improve the performance of automatic speech

recognition systems. Normalization schemes, which are introduced in detail in Chapter 2, will be the focal point of this work.

Signal analysis is an autonomous part of modern speech recognition architectures (cf. Figure 1.1). It is based on short-term spectral analysis (basic principles of spectral analysis are described in [Rabiner & Schafer 78]) and yields typically one acoustic vector every 10 milliseconds. In recent years, two different signal analysis schemes have become popular. They are based on either *Mel-frequency cepstral coefficients* (MFCC [Davis & Mermelstein 80]) or *perceptual linear prediction* (PLP [Hermansky 90]).

Throughout this work, the MFCC-based RWTH signal analysis front-end will be used, which is depicted in Figure 1.2 and described in detail in [Welling 99].

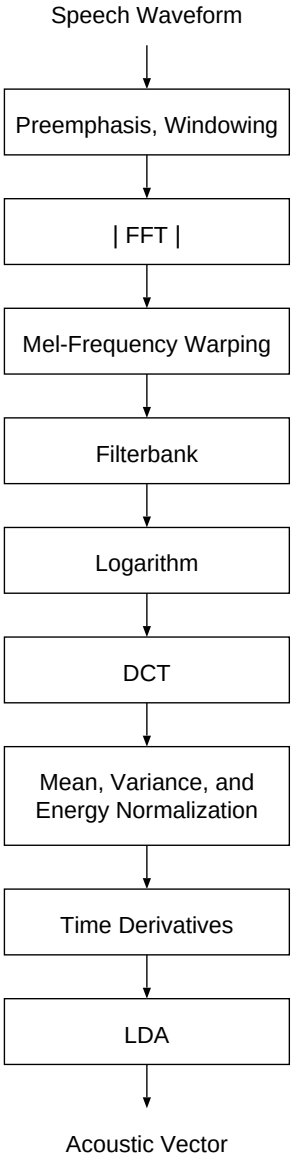

Figure 1.2: Signal analysis front-end of the RWTH speech recognition system.

First, the speech signal sampled at 8 kHz (telephone data) or 16 kHz (microphone data) is differentiated (*preemphasis*), i.e. the previous sample is subtracted from the current sample. Preemphasis is a high-pass filter that counteracts the drop in spectral energy at higher frequencies.

Every 10 ms, the preemphasized speech signal is segmented into *windows* of 25 ms length, i.e. there is an overlap of 15 ms to the left and right neighbor window. The underlying idea is that the speech signal is quasi stationary for 20 to 50 ms, which supports short-term spectral analysis within a window of 25 ms length.

Cutting a window out of the sample stream is synonymic to the application of a rectangular window function. The multiplication of speech samples with a window function in the time domain, however, is equivalent to a convolution of the speech and the window functions' spectra in the frequency domain. Hence, windowing affects the spectra derived by Fourier analysis from the speech samples. The side lobes of the spectrum of a *Hamming window* are significantly smaller than those of a rectangular window. For this reason, the Hamming function is multiplied to the windowed samples before the Fourier transform is carried out.

A number of zero samples are attached to the windowed samples (*zero padding*) to increase the number of frequency lines obtained by the subsequent fast Fourier transform (FFT). In the RWTH signal analysis front-end, a magnitude spectrum sampled at 512 discrete frequencies is derived.

In the next step, the frequency axis is warped according to the *Mel-scale* [Young 93]. As the result, the spectral resolution is reduced towards higher frequencies similar to the frequency response of the human ear.

The Mel-frequency magnitude spectrum passes a *filter bank* of typically 15 (telephone data) or 20 (microphone data) equidistant overlapping triangular bandpass filters. They reduce the spectral resolution and compress the information content into a few coefficients.

The dynamic range of the individual filter bank channels is reduced by taking the *logarithm*. One of the reasons for this step is that it mimics the non-linear dependency between the intensity of a speech signal and the loudness perceived by the human ear. Another reason is that convolutional distortions introduced to the speech signal by the transmission channel are multiplicative in the spectral domain. By taking the logarithm they become additive and can be removed easier by subsequent normalization steps.

The *discrete cosine transform* applied to the log filter bank coefficients uncorrelates the filter bank channels. The highest cepstral coefficients are typically omitted, as they contain only little information about the spoken word sequence. The resulting vector of typically 12 (telephone data) or 16 (microphone) coefficients is the standard MFCC vector.

Mel-Frequency warping, filter bank, log compression, and cepstral uncorrelation will be discussed in detail in Section 6.7 again, where an alternative method to compute

Mel-frequency cepstral coefficients directly from the magnitude spectrum is proposed.

Subsequent signal analysis steps do not belong to the "standard" MFCC signal analysis anymore, but they have shown to improve the recognition performance of automatic speech recognition systems significantly, which is why they are a fixed part of the RWTH signal analysis front-end.

The long-term mean (typically sentence mean) of each cepstral coefficient is subtracted from the MFCC vector to remove time-invariant distortions introduced by the transmission channel and the recording device (*cepstral mean normalization*, CMN). Sometimes the coefficients are also transformed to unity variance (*cepstral variance normalization*, CVN). Both methods are in fact normalization schemes, which will be discussed in detail in Section 3.1.3. Mean normalization is always applied, whereas variance normalization is only applied on corpora with large variations in the acoustic signal (cf. Chapter 5).

Energy normalization is carried out sentence-wise as well. The maximum of the zeroth cepstral coefficient within the sentence is subtracted from the zeroth coefficient of each cepstrum vector (which is proportional to the log energy of the time frame). This reduces the energy of speech frames to approximately the same level independent of the loudness. At the same time, however, a larger sentence-dependent variation in the energy level of silence is introduced.

Next, the normalized MFCC vector is augmented with *times derivatives*. Typically the first derivatives of all cepstral coefficients, and the second derivative of the zeroth cepstrum coefficient are computed by linear regression from five successive cepstrum vectors. Increasing the temporal context of Mel-frequency coefficients improves their quality.

Finally, three successive augmented Mel-frequency vectors are concatenated to yield one large vectors, which is transformed by *linear discriminant analysis* (LDA). At the same time, the size of the final acoustic vector is reduced to a desired dimension of typically 25 (telephone data) or 33 (microphone data) coefficients. LDA is a standard method in pattern recognition. It transforms feature vectors into a sub-space which maximizes the separability of different classes (e.g. phonemes). The first dimensions of the resulting LDA-transformed acoustic vector are most discriminant, which is why there is essentially no loss of information involved by the dimension reduction.

Alternatively to the time derivatives, linear discriminant analysis can be applied to a larger sequence of seven or nine normalized MFCC vectors without their derivatives. Recent experiments have shown that this method yields typically a slightly better recognition accuracy. Which method was actually applied for the different corpora used in this work is summarized in Chapter 5.

1.3 Acoustic Modeling

The aim of acoustic modeling is to provide the acoustic probability that a hypothesized word sequence W generates an observed sequence of acoustic vectors X. These probabilities are trained on large speech corpora.

Only in rare cases it is possible to train acoustic models for whole words (e.g. in the case of digit recognition). The recognition vocabulary contains typically many words that are not (or not frequently enough) observed in the training data. For this reason, acoustic modeling is typically based on sub-word units. Models for whole words are built from a concatenation of these units according to the *pronunciation lexicon*.

Phonemes are the basic sounds human speech is made of. In the RWTH system, the pronunciation lexicon is based on a language-dependent inventory of typically 40 to 50 phonemes. Depending on their context, phonemes may be articulated in a different way. Hence, the sub-word units used for acoustic modeling in most speech recognition systems are phonemes in their phonetic context (i.e. conditioned by their left and right neighbor phoneme), called *triphones*.

The same word (or phoneme) can be uttered at different speaking rates as well, so it can generate acoustic vector sequences of different length. For this reason, *hidden Markov models* [Baker 75][Rabiner 89] are used for triphone modeling in automatic speech recognition.

Hidden Markov models are stochastic finite state automata. They consist of a number of states and transitions between these. Each state is characterized by the probability to observe a given acoustic vector (*emission probability*), and the probability to step into one of the possible successor states (*transition probability*). An acoustic model θ is the sum of all hidden Markov model parameters that describe the sub-word units of a speech recognition system (Eqn. 1.4):

$$p(X|W) \quad \overset{HMM}{=} \quad p(X|W;\theta) \tag{1.4}$$

Based on the pronunciation lexicon, the word sequence W is decomposed into a sequence of triphones, which is modeled by a sequence of hidden Markov model states S.

As an example, Figure 1.3 depicts the HMM topology for a part of the word "seven" in the RWTH speech recognition system. The word is composed of the four phonemes *s*, *eh*, *v* and *un*. Each triphone (e.g. the triphone ${}_{s}eh_{v}$, which is the phoneme *eh* with the left context *s* and right context *v*) is modeled by a 6-state hidden Markov model with strict left-to-right topology. The model consists of three segments (marked <1>, <2>, and <3>), each of which is made up of two identical states [Schwartz & Chow^{+} 85][Ney & Noll 88].

To model different temporal realizations of words and their sub-word units, three types of state transitions are allowed (Bakis topology [Bakis 76]). The automaton may stay in

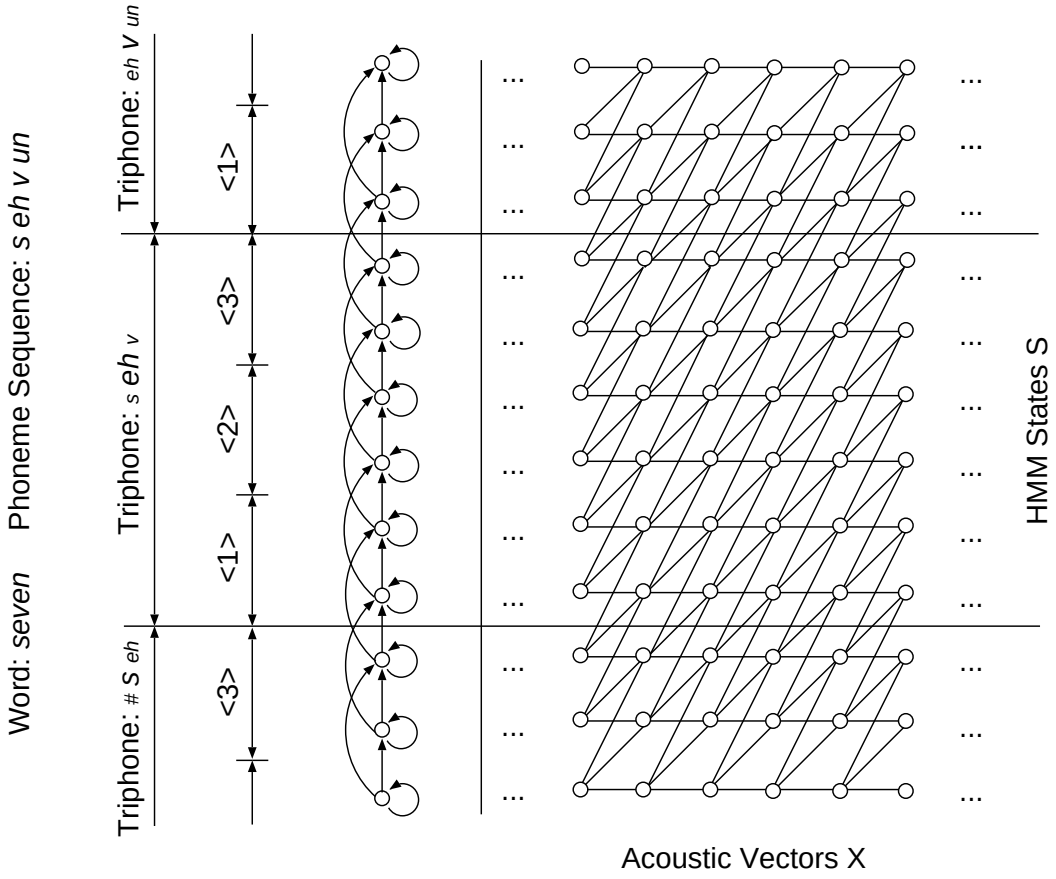

Figure 1.3: 6-state hidden Markov model in Bakis topology for the triphone ${}_{s}eh_{v}$ in the word "seven". The HMM segments are marked <1>, <2>, and <3>.

the same state (*loop*), enter the next state (*forward*), or the state after the next state (*skip*).

Under the assumption that the acoustic vector sequence follows a first order *Markov process* [van Kampen 92], the emission and transition probabilities depend only on the previous and current HMM state. Hence, the probability of an acoustic vector sequence X given the HMM state sequence S (which is derived form the word sequence W) can be expressed as follows:

$$p(X|W;\theta) = \sum_{s_1^T} \prod_{t=1}^{T} \{p(s_t|s_{t-1},W) \cdot p(x_t|s_t,W;\theta)\} \tag{1.5}$$

The sum is computed for all possible alignments s_1^T between HMM states and acoustic vectors, i.e. over all possible paths through the grid spanned by the HMM states S and acoustic vectors X in Figure 1.3. The probability of each path is given by the product over all time frames $t = 1,\ldots,T$ of the transition probability $p(s_t|s_{t-1},W)$ and the emission probability $p(x_t|s_t,W;\theta)$.

In practice, the sum over all possible paths s_1^T is often replaced by the maximum (*Viterbi* or *maximum approximation* [Ney 1990], Eqn. 1.6). This is appropriate since the feature space is of high dimension and the distribution of data is assumed to be of an exponential type. Hence, there is typically only one path that contributes the majority to the sum:

$$p(X|W;\theta) \cong \max_{s_1^T} \prod_{t=1}^{T} \{p(s_t|s_{t-1},W) \cdot p(x_t|s_t,W;\theta)\} \tag{1.6}$$

Both the summation and the maximum approximation can be efficiently calculated by the *forward-backward algorithm* [Rabiner & Juang 86] or by *dynamic programming* [Ney 84].

If only forward transitions are used, it takes six time frames to get through a hidden Markov model. Given the frame shift of 10 ms (cf. Section 1.2), this amounts to 60 ms which is approximately equivalent to the typical duration of phonemes. In tests on the VerbMobil II corpus (cf. Section 5.2.2) it was found, however, that especially for fast conversational speech the minimum duration of 30 ms (three skips) is still too large. Hence, for this corpus the two states per HMM segment are merged into a single state (3-state HMM topology). Since all three transition types (loop, forward, and skip) are still allowed, the dwell time can be reduced to only 10 or 20 ms per hidden Markov model. The 3-state HMM topology performed worse on all other corpora, though.

The emission probabilities can be modeled either as *discrete probabilities* [Jelinek 76], as *semi-continuous probabilities* [Huang & Jack 89], or as *continuous probability distributions* [Levinson & Rabiner$^+$ 83]. The latter case is applied in the RWTH speech recognition system. *Mixtures densities* made of a weighted sum of Gaussian distributions (Eqn. 1.7) are used to model the continuous probability distributions. The Viterbi approximation [Viterbi 67] is applied at the density level as well (Eqn. 1.8):

$$\begin{aligned} p(x_t|s_t,W;\theta) &= \sum_{l=1}^{L} c_{sl} \cdot \mathcal{N}(x_t|\mu_{sl},\Sigma,W;\theta) && (1.7) \\ &\cong \max_{l} \{c_{sl} \cdot \mathcal{N}(x_t|\mu_{sl},\Sigma,W;\theta)\} && (1.8) \end{aligned}$$

Here l denotes the index of the density within the mixture of state s, and c_{sl} the mixture weight. μ_{sl} is the mean vector of Gaussian density l in state s, and Σ the covariance matrix. The RWTH system uses a pooled diagonal covariance matrix, i.e. Σ is independent of s and l.

During training, each mixture is initialized by a single Gaussian density. Later the acoustic resolution is increased successively by density splitting. Parameter estimation is performed according to the *maximum likelihood* principle (Eqn. 1.9) with the *expectation maximization* algorithm [Dempster & Laird$^+$ 77], an iterative scheme that guarantees convergence to a local optimum. A more detailed description of the training procedure is given in [Beulen 99]:

$$\theta = \arg\max_{\theta'} p(X|W;\theta') \tag{1.9}$$

Among other reasons, sub-word units were introduced because not all words that are to be recognized occur in the training data. However, when context-dependent phonemes are used, there will still be a number triphones in the recognition vocabulary that are never seen in training. One solution is to fall back to context-independent monophone models, but more effective are phoneme or state clustering methods. *Decision-tree* based top-down clustering algorithms [Hwang & Huang$^+$ 92] at the HMM state level have prevailed in automatic speech recognition. The basic idea is that a large pool of generalized HMM states is created. Which of these states is tied to a specific segment of a specific triphone hidden Markov model is determined by the decision tree. Details of the state tying approach in the RWTH speech recognition system are presented in [Beulen 99].

Special care is required for triphones at word boundaries. A dummy symbol "#" is used in the case of *within-word modeling* to represent the word boundary context (cf. the first triphone in Figure 1.3). However, coarticulation across word boundaries occurs frequently in continuous speech, i.e. the last phoneme of a word is affected by the first phoneme of the successor word, and vice versa. *Across-word modeling* [Hon & Lee 91][Odell & Valtchev$^+$ 94] takes explicitly care of word boundary triphones by modeling transitions both with and without coarticulation. It results in a consistent reduction of the word error rate at the cost of a significant increase in computational complexity. Thus, special care has to be taken for an efficient implementation of across-word models. Details about across-word modeling in the RWTH system are given in [Sixtus 02]

1.4 Language Modeling

The aim of the language model is to provide the prior probability $p(W)$ of a word sequence independent of the acoustic signal. It covers syntax, semantics, and pragmatics of a language which does not mean, however, that grammatic rules are explicitly coded into the language model. Stochastic models that predict the probability of a word given the sequence of predecessor words (called the *history* of a word, Eqn. 1.10) have shown the best performance and are used in virtually every speech recognition system. To a certain degree they implicitly learn the semantic and syntactic rules of a given language.

In order to estimate the probability distribution from large text corpora (e.g. newspaper texts), the history has to be limited. Under the assumption that the word sequence follows an $(m-1)$-order Markov process [van Kampen 92], the probability of a word depends only on the $(m-1)$ predecessor words (Eqn. 1.11), and the corresponding models are called *m-gram language models* [Bahl & Jelinek$^+$ 83]:

$$p(w_1^N) = \prod_{n=1}^{N} p(w_n|w_1^{n-1}) \tag{1.10}$$

$$\cong \prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1}) \tag{1.11}$$

If the lower index $n-m+1$ is smaller than one it is set to one with $p(w_1|w_1^0) = p(w_1)$.

As for training of acoustic models, the maximum likelihood principle is also applied for training of the language model with the *perplexity* (PP [Bahl & Jelinek$^+$ 83]) as the evaluation criterion. The perplexity of a word sequence is defined as the inverse geometric mean of the language model probability over all words in the sequence (Eqn. 1.12). The perplexity of a test utterance can be viewed as the average number of possible successor words at each instant in the search process, which is typically well below the vocabulary size:

$$PP(w_1^N) = \left\{ \sqrt[N]{\prod_{n=1}^{N} p(w_n)} \right\}^{-1} \tag{1.12}$$

Using the perplexity as training criterion that is to be minimized over the text database yields the solution that the probability of an m-gram is given by its relative frequency in the training corpus. However, with increasing history length the number of m-grams increases exponentially, and even for *trigram language models* that are used in most speech recognition systems ($m = 3$, i.e. the probability of each word is conditioned by the two predecessor words) and training corpora of several million running words, the vast majority of trigrams will not be seen in the data or occur too infrequent. Unseen m-grams would get the probability zero and could never be recognized, which is why smoothing needs to be applied to ensure that the probability of all m-grams is larger than zero.

Smoothing methods are based on various *discounting* schemes [Katz 87][Ney & Essen$^+$ 94] that reduce the probability mass of observed m-grams to distribute it among the unseen (*backing off*) or all (*interpolation*) m-grams. The amount of discounting mass that is assigned to each m-gram is typically governed by generalized language model probabilities based on shorter histories. The *leaving-one-out* algorithm is the method of choice to estimate the discounting and generalized language model parameters. A systematic comparison of smoothing techniques based on the RWTH speech recognition system is given in [Martin & Hamacher$^+$ 99].

A number of schemes were proposed to increase the performance of language models. The idea of a *language model cache* [Kuhn & de Mori 90] is to use the last few hundred recognized words for adaptation of the language model to the current topic of conversation. Sequences of words that occur frequently in the same order are pooled to *phrases*

which are handled as a single words and effectively increase the history of the language model [Jelinek 91]. Similar words (e.g. proper names) can be pooled to *word classes* [Brown & Della Pietra$^+$ 92] for more robust parameter estimation, and *distant m-grams* [Rosenfeld 94] are m-grams with a gap between the current word and the $(m-1)$ predecessor words that condition it.

A description of the RWTH language model implementation is given in [Wessel & Ortmanns$^+$ 97]. The language models used for recognition tests reported in this work can be characterized as:

- trigram language models
- smoothing is achieved by absolute discounting with interpolation
- smoothing parameters are estimated by leaving-one-out
- the discounted probability is distributed among all trigrams based on *singleton generalized backing-off distributions*
- word classes are used for some corpora (cf. Chapter 5)
- phrase language models are used for some corpora, but not those employed in this work; cache and distant m-gram techniques are not applied

1.5 Search

Based on the two knowledge sources acoustic model and language model, the task of the search is to find the best word sequence. The optimization criterion is the posterior probability (cf. Section 1.1), which is proportional to the product of the acoustic and the language model probability (cf. Eqn. 1.3). Replacing the two terms according to Eqn. 1.6 and 1.11 yields:

$$W = \arg\max_{W'} \left\{ \prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1}) \cdot \max_{s_1^T} \{ \prod_{t=1}^{T} p(s_t|s_{t-1}, W') \cdot p(x_t|s_t, W'; \theta) \} \right\} \quad (1.13)$$

The search has (at least in theory) to hypothesize all possible word sequences $W = w_1, \ldots, w_N$ and find the one with maximum likelihood according to Eqn. 1.13. A naive implementation that computes the probability of all word sequences in infeasible, since for a given vocabulary size V the number of possible word sequences grows exponentially with the number N of words in the sequence:

$$V^0 + V^1 + V^2 + V^3 + \ldots + V^N = \frac{V^{N+1} - 1}{V - 1} \quad (1.14)$$

The complexity of the optimization can be reduced significantly by dynamic programming [Bellman 57], which exploits the mathematical structure of the task and decomposes the

global into a number of local optimization problems. Two search algorithms based on dynamic programming have become popular in automatic speech recognition.

In *A*-search* or *stack decoding* [Jelinek 69], a time-asynchronous expansion of state hypotheses is applied which relies on a heuristic estimate of the probability of the remaining unexpanded path. Convergence to the global optimum is guaranteed if the estimate is strictly above the true probability. The efficiency of A*-search heavily depends on the quality of the heuristic estimate.

In the case of *Viterbi search* [Vintsyuk 71][Ney 84] which is applied in the RWTH system, state hypotheses are expanded time-synchronous. The advantage is that at each time frame the likelihood of all hypotheses can be compared with each other, which allows for efficient *pruning techniques*. Unlikely hypotheses are omitted early from the optimization process, which significantly reduces the search space.

A number of methods are applied to further reduce the effort of finding the best word sequence:

- The pronunciation lexicon is organized as a *prefix tree* [Ney & Haeb-Umbach$^+$ 92], which exploits redundancies in the lexicon and reduces the search space.

- Pruning is applied when states are expanded to pursue only the most promising hypotheses (*beam search* [Ney & Mergel$^+$ 87][Ortmanns & Ney 1995]). It may happen that the globally best word hypothesis is not found, since it could be pruned beforehand due to poor likelihood at an intermediate search stage. However, a proper adjustment of pruning parameters ensures that no significant search errors occur.

- A number of *look-ahead techniques* are applied to make pruning more efficient. If the pronunciation lexicon is organized as a prefix tree, the identity of the hypothesized word is not known until the end of the tree is reached, which is why the language model probabilities can be introduced only at word ends. However, at each node within the tree the set of possible ending words is limited. Hence, upper estimates for the language model probability can be propagated backwards into the tree nodes and integrated earlier into the search process (*language model look-ahead* [Steinbiss & Tran$^+$ 94]). Furthermore, the approximate acoustic probability of the next few acoustic vectors can be estimated in advance using simplified acoustic models, and subsequently integrated into the search process (*phoneme look-ahead* [Ney & Haeb-Umbach$^+$ 92]).

- A large fraction of the computation time in automatic speech recognition systems is spend for the calculation of the acoustic emission probabilities (cf. Eqn. 1.8). A number of *fast likelihood calculation* techniques were proposed to reduce this effort. Examples are algorithms to structure the search space [Fritsch 97], to quantize the acoustic vectors [Bocchieri 93], or to partition the acoustic feature space [Nene & Nayar 1996]. A comprehensive overview of these techniques and their implementation in the RWTH system is given in [Ortmanns & Ney$^+$ 97] and [Ortmanns 1998]. In addition, a significant reduction in computation time is

achieved by parallelized likelihood calculation based on SIMD (single instruction, multiple data) instructions of the microprocessor [Kanthak & Schütz$^+$ 00].

The idea of *multi-pass* as opposed to *integrated* or *single-pass search* is that a simplified acoustic and/or language model is used in a first recognition pass. Not only the first best word sequence, but also competing hypotheses with a similar likelihood are saved in either a *N-best list* or a *word graph* for further processing. In a N-best list, the N word sequences with highest posterior probability are kept [Schwartz & Chow 90]. A word graph is a directed acyclic graph that stores intermediate word sequences as arcs [Schwartz & Austin 91]. In further recognition passes, more elaborate acoustic and/or language models are used to refine the likelihood of the hypotheses at comparatively low computational costs, since the search space is largely restricted.

For recognition tests reported in this work, single-pass beam search with conservative pruning settings and language model look-ahead was applied. Parallelized fast likelihood calculations were used in all tests, but the other techniques listed above were only applied in one test were the recognizer was accelerated to almost real-time (Section 6.3.4).

Chapter 2

Adaptive Acoustic Modeling

2.1 Introduction

The acoustic signal contains a lot of variability. On the one hand this is necessary to discriminate between different speech units (e.g. phonemes), but on the other hand there are also variations in the speech signal which are irrelevant for the recognition process. Sources of irrelevant variability are, for example [Sankar & Lee 95]:

- varying transducers and transmission channels
- different speakers, speaking styles, or accents
- a varying ambient or channel noise

The training data contain typically a number of different acoustic conditions (i.e. different speakers, speaking styles, transmission channels, etc.). A particular test utterance, on the other hand, has usually only one specific acoustic condition (i.e. a specific speaker with a particular accent and vocal tract length, a specific transmission channel, etc.) The test condition may or may not be present in the training data, but it will usually differ from the mixture of training conditions. This fundamental mismatch between training and test causes degraded performance of automatic speech recognition systems:

- speaker-dependent systems trained on data from the test speaker significantly outperform speaker-independent systems trained on data from different speakers
- gender-dependent systems trained on male or female data only perform usually better on corresponding test speakers than gender-independent systems
- systems trained on very fast/slow speech outperform universal systems trained on data with different speech rates in the case of very fast/slow test speakers
- a system that was trained on data collected in a car or over a telephone channel gives superior performance in the same test environment compared to a system trained on studio quality microphone data

A schematic view of training and test is shown in Figure 2.1. The left side depicts the feature space, i.e. the sequence of acoustic vectors X and its generation, and the right side shows the model space, i.e. the acoustic model θ and its training.

In general, three abstract data levels can be distinguished:

- the training data (first level), which are a collection of different conditions
- a test utterance (third level), which is usually of one specific condition only
- the intermediate reference level (second level), at which the variations caused by different conditions are ideally removed (e.g. vocal tract length normalized or speaker-adapted acoustic data and models)

In this framework, speech recognition can be regarded as a combination of acoustic vectors and an acoustic model from specific data levels. There is a mismatch if they do not belong to the same level. In the case of conventional non-adaptive acoustic modeling, for example, there is a strong mismatch between test data X_{Test} and the acoustic model θ_{Train}.

Two basic strategies of adaptive acoustic modeling, which are depicted in Figure 2.2, may be employed to avoid the mismatch: Either condition-dependent acoustic models θ_{Test} are trained such that there is a-priori no mismatch, or the mismatch is removed by transformation of the acoustic vectors and/or the acoustic model onto a different level.

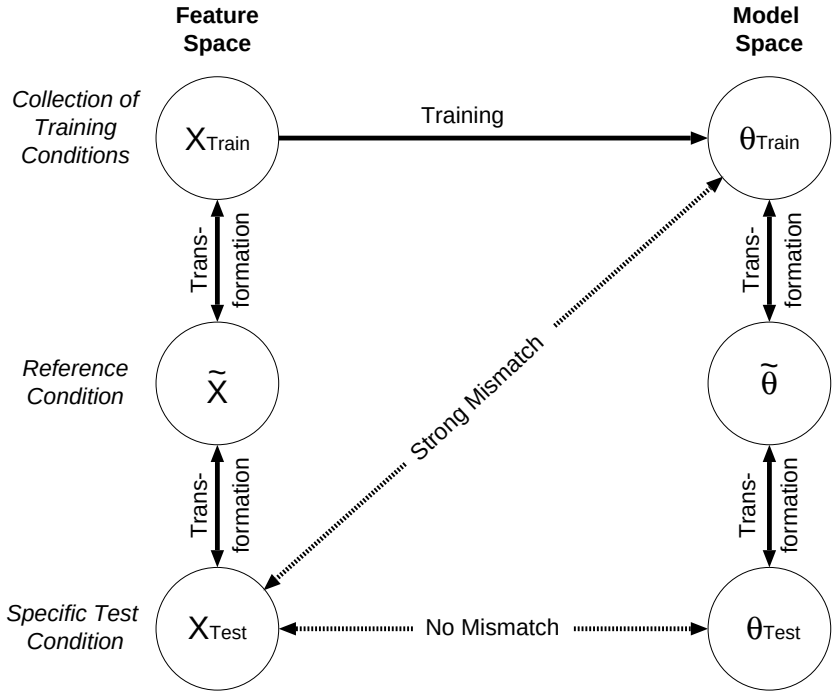

Figure 2.1: Schematic view of training and test. Depicted are the feature and the model space as well as three abstract data levels.

In the first case (depicted on the left side of Figure 2.2), specialized acoustic models are trained for each possible test condition (e.g. speaker-dependent or gender-dependent models, models for fast, average and slow speech, models for microphone and telephone data, etc.). Since the condition of a particular test utterance is a-priori unknown it is not clear, however, which of the acoustic models has to be used to recognize the utterance. Two possible solutions are:

- An own recognition pass is carried out with each specialized acoustic model. Later the system output is either generated from a combination of the individual recognition results, or by choosing the result with maximum likelihood. Since for each possible condition and own recognition pass is required, the computational load increases dramatically, which is why this approach is usually prohibitive.

- The current test condition is determined first (e.g. by means of a speaker or gender recognition system, a classifier for the rate of speech or bandwidth, etc.), and the corresponding acoustic model is selected for recognition. This approach is computationally attractive, as usually only a single recognition pass is required.

In general, the first strategy to use specialized acoustic models has one advantage, but also some major disadvantages:

Advantage:

- if the test condition occurred frequently in the training data, the acoustic model matches perfectly to the test condition

Disadvantages:

- a set of discrete conditions has to be defined (e.g. slow, average, and fast speech), and the training and test utterances have to be assigned to one of these

- the training database is split into smaller pieces corresponding to each condition, hence, there is less training data available for each specialized acoustic model and parameter estimation is less robust

- only conditions observed in training can be recognized well

The second strategy (depicted on the right side of Figure 2.2) requires only one universal acoustic model that is trained on all data. The mismatch is treated explicitly in training and test by transformation of the acoustic vectors (*normalization* or *feature transformation*), i.e. by reducing the variability of the speech signal during signal analysis, or by transformation of the acoustic model (*adaptation* or *model transformation*), which amounts to adapting the acoustic model to a specific test condition.

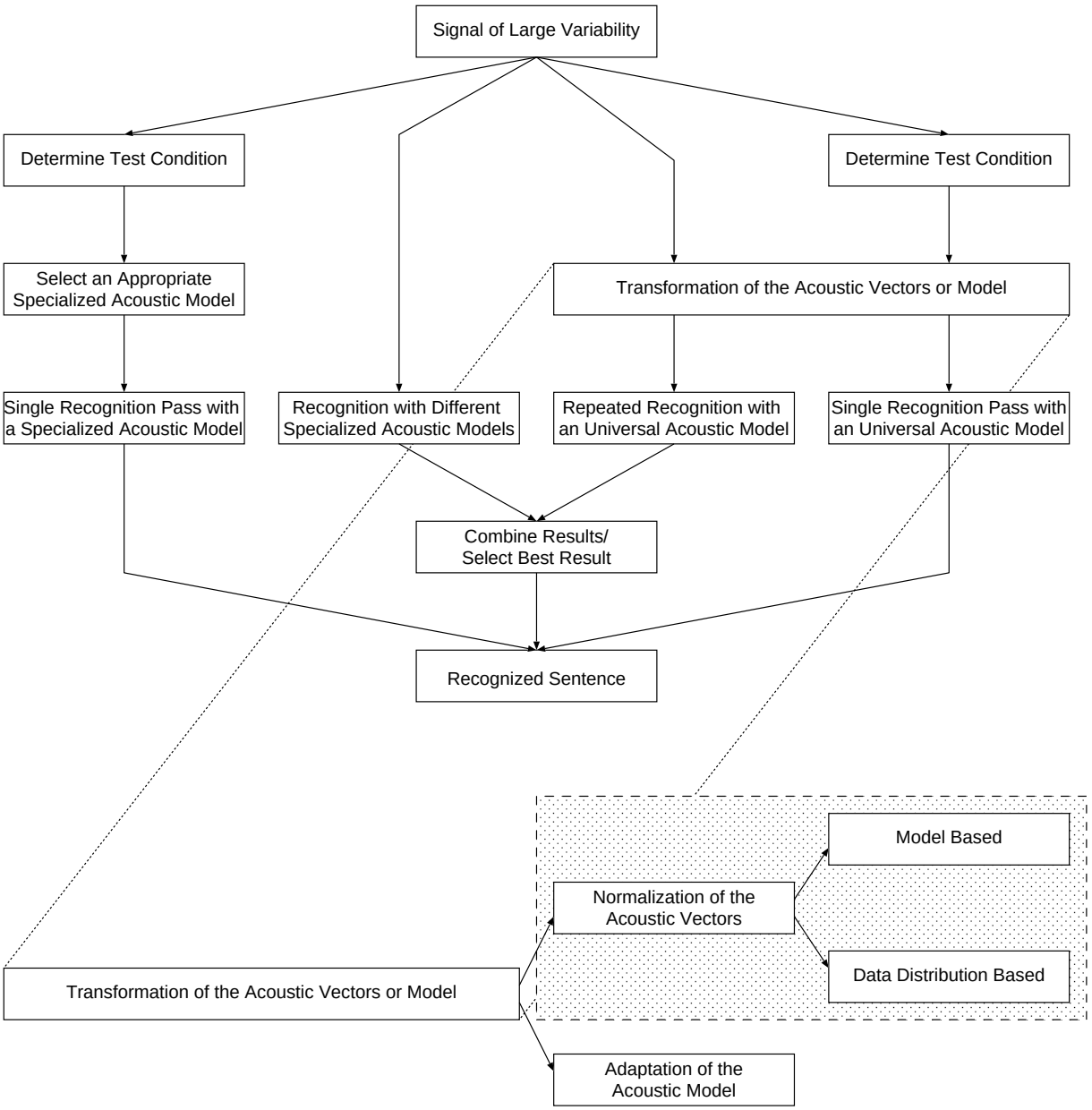

Figure 2.2: Classification of different adaptive recognition schemes. The shaded box marks those techniques which will be covered in detail in this work.

Also in this case, the condition of a particular test utterance is a-priori unknown, and the same two solutions as for specialized acoustic models are applicable: Either the transformation is repeated for each possible test condition, and own recognition pass is carried out, and the final recognition output is generated from a combination of the individual recognition results. Alternatively, the test condition is determined first, and the acoustic vectors/model are transformed accordingly before recognition.

Transformation based adaptive modeling has certain advantages:

- the acoustic model is trained on the full training database which results in more robust parameter estimation
- transformations allow for continuous estimates of the test condition (e.g. the specific rate of speech measured in phonemes per second) and corresponding normalization or adaptation
- good results are also achieved for test conditions that were not observed in training

Two disadvantages are:

- suitable transformation functions have to be found that allow for efficient compensation of a certain condition
- much adaptation data is sometimes required to achieve the performance of a specialized acoustic model

Normalization and adaption in training and test will be discussed in the following section. Afterwards, a detailed mathematical formulation of normalization and adaptation is presented. It will show how adaptive acoustic modeling fits into the framework of statistical speech recognition.

2.2 Normalization and Adaptation

In Figure 2.3, the schematic view of training and test is depicted again, this time showing the different transformations applied in practice in normalization and adaptation.

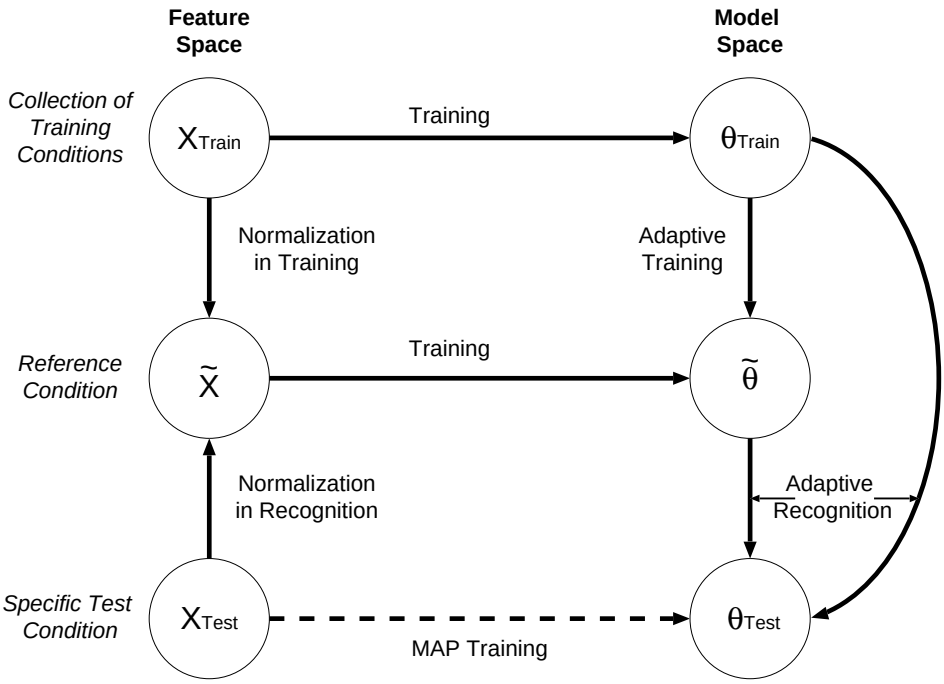

Figure 2.3: Overview of normalization and adaptation concepts.

As explained in the previous section, the mismatch between test data X_{Test} and the acoustic model θ_{Train} in the case of non-adaptive acoustic modeling is strong. Normalization transforms the acoustic vectors to a different level, whereas adaptation amounts to transforming the acoustic model.

Adaptation schemes (e.g. *maximum likelihood linear regression* [Leggetter & Woodland 95]) are capable to adapt an acoustic model trained on different conditions directly to one specific test condition ($\theta_{Train} \rightarrow \theta_{Test}$). For this reason, adaptation is usually successful even when carried out in test only.

Normalization of acoustic vectors (e.g. *vocal tract length normalization* [Lee & Rose 96]) results in a transformation into the reference condition. For this reason, there is often a moderate gain in recognition accuracy if normalization is applied in test only, since a minor mismatch between $\tilde{X}$ and θ_{Train} remains. The best performance is typically achieved if both training and test data are normalized (no mismatch between $\tilde{X}$ and $\tilde{\theta}$).

Adaptation or normalization in training alone is counterproductive. In this case, the acoustic model is adapted to a reference condition, but cannot cope well with test conditions that deviate from the average (mismatch between X_{Test} and $\tilde{\theta}$, e.g. [Welling & Kanthak$^+$ 99][Gales 01]).

2.3 Mathematical Framework

As was shown in Chapter 1, the statistical approach to automatic speech recognition amounts to finding the most probable word sequence W, i.e. the one that maximizes the product of the language model probability $p(W)$ and the acoustic probability $p(X|W;\theta)$ (cf. Eqn. 1.3 and 1.4). X denotes the sequence of acoustic vectors and θ is the acoustic model:

$$W = \arg\max_{W'} \{p(W') \cdot p(X|W';\theta)\} \tag{2.1}$$

The acoustic probability is typically modeled with first order hidden Markov models. In the Viterbi approximation, the maximum over all possible alignments s_1^T between HMM states S and acoustic vectors X is used (cf. Eqn. 1.6). The probability of each alignment is given by the product over all time frames of the transition and emission probability:

$$p(X|W;\theta) \;\cong\; \max_{s_1^T} \prod_{t=1}^{T} \{p(s_t|s_{t-1},W) \cdot p(x_t|s_t,W;\theta)\} \tag{2.2}$$

Equations 2.1 and 2.2 are employed in conventional non-adaptive modeling where no distinction is made under which condition the acoustic signal was recorded. In the previous sections it was shown, however, that the acoustic data from training and test do not match. They usually originate from different conditions (different speakers, speaking

styles, transmission channels, etc.) which can be expressed mathematically by a new condition-dependent parameter α (Eqn. 2.3). For simplicity it is assumed that only the emission probabilities are affected by the variable recording condition (Eqn. 2.4):

$$\begin{aligned} p(X|W;\theta) &\rightarrow p(X|W;\theta,\alpha) && (2.3)\\ &\cong \max_{s_1^T} \prod_{t=1}^{T} \{p(s_t|s_{t-1},W)\cdot p(x_t|s_t,W;\theta,\alpha)\} && (2.4)\end{aligned}$$

To handle the unknown parameter, it is treated as a continuous valued hidden variable that has to be integrated out (Eqn. 2.5). To avoid integration problems, the maximum approximation is applied at this stage as well (Eqn. 2.6):

$$\begin{aligned} p(X|W;\theta) &= \int d\alpha\; p(X,\alpha|W;\theta)\\ &= \int d\alpha\; p(\alpha|W;\theta)\cdot p(X|W;\theta,\alpha) && (2.5)\\ &\cong \max_{\alpha} \{p(\alpha|W;\theta)\cdot p(X|W;\theta,\alpha)\} && (2.6)\end{aligned}$$

Thus, for adaptive modeling Bayes' decision rule in the maximum approximation can be rewritten as:

$$W \cong \arg\max_{W'} \Big\{ p(W') \cdot \max_{\alpha} \{p(\alpha|W';\theta)\cdot p(X|W';\theta,\alpha)\} \Big\} \qquad (2.7)$$

The prior distribution $p(\alpha|W;\theta)$ is often assumed to be uniform.

The training corpus contains a number of conditions $r = 1,\ldots,R$. For each training condition, acoustic data X_r along with the transcriptions W_r are given. Since the condition dependent parameter α_r is unknown, adaptive training of a normalized acoustic model (cf. Eqn. 1.9) becomes a complex optimization problem:

$$\tilde{\theta} \cong \arg\max_{\theta} \prod_{r=1}^{R} \max_{\alpha} \{p(\alpha|W_r;\theta)\cdot p(X_r|W_r;\theta,\alpha)\} \qquad (2.8)$$

In practice, the parameter $\hat{\alpha}_r$ of each condition is often estimated beforehand by some function $h(\cdot)$. In the simplest case (e.g. in the case of histogram normalization [Dharanipragada & Padmanabhan 00]), $\hat{\alpha}_r$ depends on the acoustic data only (Eqn. 2.9). However, $h(\cdot)$ may also be text-dependent, and it may require an own acoustic model (e.g. in the case of vocal tract length normalization [Lee & Rose 96]):

$$\hat{\alpha}_r = h(X_r) \qquad (2.9)$$

When the parameters $\hat{\alpha}_r$ are estimated, the training data are normalized and the acoustic model $\tilde{\theta}$ is trained as usual by maximum likelihood:

$$\tilde{\theta} = \arg\max_{\theta} \prod_{r=1}^{R} p(X_r|W_r;\theta,\hat{\alpha}_r) \tag{2.10}$$

In practice, the dependency of the acoustic model from the condition r has to be defined. Typically is implemented by transformations, whereby the functional form of the acoustic model is usually fixed and only the transformation parameters are estimated from data. As shown in the previous section, there are two possible realizations to match different conditions and derive adapted probabilities $\tilde{p}(\cdot)$:

In normalization, the transformation $f_\alpha(\cdot)$ is applied to the acoustic vectors:

$$X \rightarrow \tilde{X} = f_\alpha(X) \tag{2.11}$$

$$\begin{aligned} p(X|W;\tilde{\theta},\alpha) &= \tilde{p}(f_\alpha(X)|W;\tilde{\theta}) \cdot \det\left(\frac{df_\alpha(X)}{dX}\right) \\ &= \tilde{p}(\tilde{X}|W;\tilde{\theta}) \cdot \det\left(\frac{d\tilde{X}}{dX}\right) \end{aligned} \tag{2.12}$$

Here, the Jacobian determinant of the transformation is included. However, in a classification task the Jacobian determinant can be omitted in some cases because the transformation is assumed to be independent of the word sequence W. In other cases (e.g. in vocal tract length normalization) it is usually assumed to be irrelevant.

In adaptation, the inverse transformation is applied to the acoustic model (Eqn. 2.13). For notational simplicity, the symbol $f_\alpha^{-1}(\cdot)$ is used for the inverse transformation:

$$\tilde{\theta} \rightarrow \theta = f_\alpha^{-1}(\tilde{\theta}) \tag{2.13}$$

$$\begin{aligned} p(X|W;\tilde{\theta},\alpha) &= \tilde{p}(X|W;f_\alpha^{-1}(\tilde{\theta})) \\ &= \tilde{p}(X|W;\theta) \end{aligned} \tag{2.14}$$

Even though adaptation and normalization are equivalent in this framework, both techniques are relevant in practice. The main challenge of adaptive modeling is to find suitable transformation functions $f_\alpha(\cdot)$ or $f_\alpha^{-1}(\cdot)$ that can compensate for the effects of a certain condition, and to estimate their parameters reliably on adaptation data. In some cases, the transformation function $f_\alpha(\cdot)$ has a simple functional form and allows for efficient parameter estimation (e.g. spectral warping as in vocal tract length normalization, which depends on a single parameter to be estimated from data), whereas the corresponding inverse transformation function $f_\alpha^{-1}(\cdot)$ for acoustic model adaptation cannot be derived easily or is much more complex (and vice versa).

The focus of this work will be on normalization techniques, i.e. during signal analysis the acoustic vectors are transformed into the reference form. Existing and novel transformations and parameter estimation methods will be proposed and evaluated.

2.4 Classification of Normalization Techniques

There are different ways to class normalization techniques in more detail. Based on whether they are derived from some physical model we can distinguish between *model based* and *data distribution based* normalization.

There are a number of environmental and speaker-dependent variations whose impacts on the speech signal are to some extent predictable. Model based normalization tries to account for such variability. It is based on some model for speech production, transmission, or perception. A small number of model parameters are estimated on adaptation data and applied according to the underlying model to account for the undesired variability. Well-known speaker normalization techniques like vocal tract [Lee & Rose 96] and speaking rate normalization [Mirghafori & Fosler$^+$ 95] are model based approaches. Other algorithms that fall into this category are channel normalization schemes like cepstral mean normalization and a number of noise suppression techniques.

If the effect of the environment is not predictable or too complex, normalization techniques that are independent of any model for speech production, transmission, or perception may be applied. These data distribution based techniques aim at transforming the acoustic vectors into a domain that is more suitable for automatic speech recognition. The transformation parameters are obtained from the distribution of the training and test data. Examples are feature space transformations like the Gaussianization technique [Gopinath 00] or feature space matching like stochastic matching [Sankar & Lee 95] and histogram normalization [Dharanipragada & Padmanabhan 00].

In Chapter 3, an overview of these and other normalization techniques proposed in the literature will be given. The published results will be discussed and open questions to be addressed in this work will be worked out in Chapter 4.

2.5 Normalization and Signal Analysis

Normalization amounts to a transformation of the acoustic vector. Typically it is not the fully processed acoustic vector that is transformed, but in most cases normalization is carried out at intermediate stages of signal analysis. Normalization may either be achieved by adapting parameters of existing signal analysis components (e.g. by modifying the frame shift to normalize the speaking rate) or by introducing additional components (e.g. an additional spectral warping step for vocal tract length normalization).

Figure 2.4 shows a typical signal analysis front-end, similar to the one used at RWTH (cf. Figure 1.2). A number of normalization techniques and the stage where they are applied during signal analysis are listed.

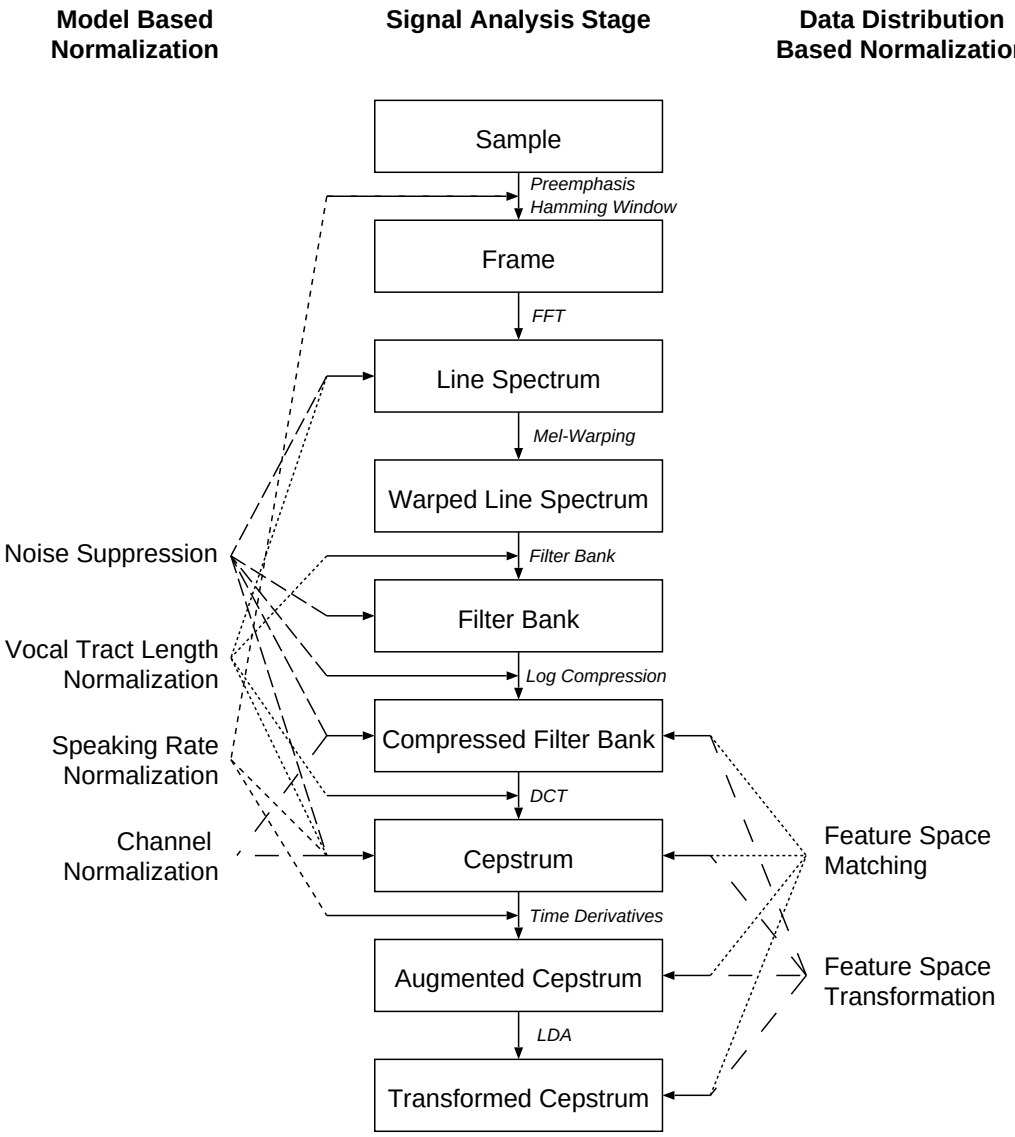

Figure 2.4: Overview of signal analysis components and stages were normalization may be applied.

A speech frame is first obtained by applying a Hamming window to a sequence of preemphasized speech samples. The frame shift, i.e. the time between successive acoustic vectors, may be modified to normalize the speaking rate.

A line spectrum is obtained from the speech frame by means of a Fourier transform. Vocal tract length normalization and some noise suppression techniques transform the line spectrum. Alternatively, these techniques may also be applied to the filter bank coefficients.

The line spectrum is warped according to the Mel-scale to account for the reduced spectral resolution of the human ear towards higher frequencies. Afterwards the warped spectrum passes a bank of overlapping triangular bandpass filters. VTN may

be implemented by a modification of the center frequencies, and all spectral warping functions (Mel-scale, VTN) can be integrated into the cepstrum transformation of the log-magnitude spectrum as will be described in Section 6.7.

The dynamic range of the filter bank coefficients is typically compressed with the logarithm or a similar function. At this stage, noise suppression and feature space matching techniques have been applied successfully.

The discrete cosine transform is applied to uncorrelate the filter bank channels. Further spectral smoothing is achieved by omitting the highest cepstral coefficients. Channel normalization is usually applied at the log filter bank or cepstrum, as a convolutional disturbance to the speech signal is multiplicative in the spectral domain, but additive in the cepstral domain. Vocal tract length normalization and noise suppression can be implemented here as well.

Finally, the cepstrum vector is augmented with time derivatives and optionally further transformed by linear discriminant analysis. Speaking rate normalization can be achieved by interpolation at the cepstrum stage or by modified calculation of the time derivatives. A number of feature space matching and transformation schemes are applied to the final acoustic vectors, which could be either the (augmented) cepstrum or the LDA-transformed vector.

Chapter 3

Normalization: State of the Art

Normalization techniques have been applied for a long time in automatic speech recognition. Some of these (e.g. cepstral mean normalization) have become a standard element of modern speech recognition front-ends which blurred the borderline between signal analysis components and additional normalization components (cf. Section 1.2). Furthermore, the variety of normalization techniques published in the literature is so large that a comprehensive overview is hard to give. For these reasons, only the most common techniques and those related to research results presented in this work will be summarized in this chapter. According to the classification introduced in Section 2.4, normalization techniques will be divided into model based and data distribution based schemes.

3.1 Model Based Normalization Schemes

3.1.1 Vocal Tract Length Normalization

In 1977, Wakita proposed a normalization scheme based on a frequency axis warping during signal analysis [Wakita 77]. The idea was to remove the shift in formant frequencies caused by different lengths of the speakers' vocal tracts. He applied VTN for the improved recognition of isolated vowels. The idea was later revivided by Kamm et al. [Kamm & Andreou$^+$ 95] during a summer workshop at Johns Hopkins University, which triggered new research in this field.

Most papers published subsequently about vocal tract length normalization addressed one or more of the following topics:

- type of the frequency axis warping function (linear, non-linear) and its implementation (time domain, frequency domain, cepstral domain)
- reliable estimation of the warping factors in training
- efficient warping factor estimation in test (with respect to word error rate, required adaptation data, and computational overhead)
- gain in recognition accuracy achieved by VTN under different conditions (clean vs. noisy environment, small vs. large training corpora, small vs. large vocabulary)

- comparison of VTN with adaptation techniques, sequential application of VTN and adaption schemes (e.g. MLLR)

In the following, a number of publications that are of particular interest will be discussed in detail.

Acero and Stern proposed a bilinear transformation in the cepstrum domain already in 1991 that resulted in non-linear warping of the frequency axis [Acero & Stern 91]. Even though they did not call it vocal tract length normalization, it was essentially the same. Acero and Stern found a clear distinction in warping factors between male and female speakers, and they achieved a 10% relative reduction in word error rate on the Census database.

In 1996, Eide and Gish investigated the impact of different frequency warping functions and of the amount of training data on the recognition performance with VTN [Eide & Gish 96]. Warping factors were estimated speaker-wise using the median position of the third formant. Best performance was reported with a non-linear warping function, but the differences to linear warping were small. On the SwitchBoard corpus, a reduction in word error rate of 8% relative was obtain when 5 hours of training data were used, which reduced to 6% relative when the full training corpus of 63 hours was utilized. Eide et al. investigated also methods to enrich the training corpus with additional normalized data, but could not reduce the word error rate any further.

At the same conference, Lee and Rose presented a paper on VTN with a number of new and successful ideas [Lee & Rose 96]. They estimated warping factors in a maximum likelihood framework. For training speakers, an iterative procedure was proposed, whereby an acoustic model was trained on one half of the normalized training data, which was then used to estimate warping factors for the other half. Subsequently the data sets were swapped and the warping factors for the first half were re-estimated with a new acoustic model trained on the second half of data. It was found that more than one iteration reduced the word error rate on the training data, but not anymore on the test data.
For test data, Lee and Rose proposed a Gaussian mixture model (GMM) based warping factor estimation. For each warping factor, a GMM was trained on all unnormalized training data with that warping factor. Each test utterance was scored with all GMMs and the warping factor corresponding to the model with maximum likelihood was used for recognition. On a telephone based connected digit recognition task, Lee and Rose achieved a reduction of word error rate of 15% relative. That compared to a 20% reduction by the baseline two-pass VTN approach, where the transcription of a first recognition pass was used for text-dependent warping factor estimation.
VTN in training and test performed better than gender-dependent modeling, since more training data could be utilized (cf. Section 2.1). Instead of re-sampling the speech waveform in the time domain, Lee and Rose proposed furthermore to incorporate linear frequency axis warping into Mel-frequency warping by modifying the center frequencies and bandwidths of the filter bank channels.

Wegmann et al. obtained warped spectra by linear interpolation from the original discrete-frequency spectrum [Wegmann & McAllaster+ 96]. They applied a piece-wise linear warping function and proposed a fast warping factor scheme similar to the one of Lee and Rose by training one generic model of normalized speech. This model was trained in an iterative fashion. The current generic speech model was used to estimate new warping factors, and on the normalized data a new generic model was trained. Only voiced frames were used for warping factor determination. On the SwitchBoard corpus, the word error rate could be reduced by 12% relative for gender-independent and by 6% relative for gender-dependent acoustic modeling.

In 1997, Zhan and Westphal compared warping factor estimation based on the median position of the first three formants with maximum likelihood estimates [Zhan & Westphal 97]. They found that the latter approach consistently outperformed formant-based estimates. A piecewise-linear warping function yielded better results than the non-linear function proposed by Eide and Gish.
To accelerate the grid search over all warping factors in two-pass VTN, Zhan and Westphal proposed to keep the alignment between acoustic vectors and HMM states from the first recognition pass fixed when performing the grid search for the best warping factor. In the optimal setup, the word error rate could be reduced by 9% relative on a 5k-word vocabulary Spanish spontaneous speech scheduling task.

Gouvêa and Stern proposed an enhanced scheme for warping factor estimation based on the median frequency of the first three formants [Gouvêa & Stern 97]. They fitted a linear warping function that did not necessarily had to insect the origin. In return they got consistently better results in clean and noisy conditions with word error rate reduction of up to 15% on the Resource Management database. At least five sentences were required to estimate a warping factor reliably. A data-driven non-linear warping function gave an additional minor improvement in recognition accuracy.

An approach for warping factor estimation based on the pitch was proposed by Chu et al. [Chu & Jie+ 97]. Their approach was slightly inferior to maximum likelihood estimates with respect to the recognition accuracy, but saved computation time. It was shown that spectral warping by modifying the filter bank was more robust than re-sampling in the time domain. Under mismatch conditions (training on male or female speakers only, test on both genders) more than 30% relative reduction of error rate was obtained on a Mandarin digit recognition task.

Pye and Woodland studied the combination of VTN and MLLR on clean large vocabulary corpora, namely different Wall Street Journal test sets [Pye & Woodland 97]. They applied spectral warping by adapting the filter bank center frequencies and did a grid search for the warping factor based on the transcription from a first recognition pass (two-pass VTN). It was found that MLLR gave typically a somewhat larger gain in recognition performance than VTN, but the gain of both techniques was to a large extent additive. They also confirmed that the gain by VTN reduced when more training data and a larger vocabulary were used. With the 15 hour WSJ0 training corpus, Pye and Woodland achieved reductions between 12% and 15% relative on two 5k and

20k-word vocabulary test sets. With the 66 hour WSJ0+1 training database, however, the improvement reduced to between 7% and 8% on the 64k-word vocabulary test set.

McDonough et al. proposed an extension to VTN called all-pass transform [McDonough & Byrne$^+$ 98]. Based on a bi-linear frequency axis warping function they showed that VTN can be expressed as a linear transformation in the cepstrum domain. A major advantage was that the Jacobian determinant of the transformation could be taken into account (cf. Section 2.3) to keep the probability distributions normalized. Recognition tests on the SwitchBoard corpus yielded reductions in word error rate by 7% relative for the bilinear transform, and 8% relative for the more general all-pass transform.

The VTN setup of RWTH was presented by Welling et al. in the same year [Welling & Haeb-Umbach$^+$ 98]. They investigated VTN and MLLR on the Wall Street Journal corpus with a 5k-word vocabulary test set. With two-pass recognition, the word error rate was reduced by 11% relative in gender-independent, and by 4% relative in gender-dependent mode. On the German SieTill database consisting of connected digit strings, Welling et al confirmed that the gain of VTN increased when simple acoustic models were used. Finally they proposed an alternative scheme for fast warping factor estimation in test based on one Gaussian mixture model for normalized speech similar to the technique proposed by Wegmann et al. On the WSJ corpus, this approach performed almost as good as two-pass VTN.

Further results with fast warping factor estimation were published by Welling et al. in 1999 [Welling & Kanthak$^+$ 99]. They proposed a simple method to omit silence frames from the warping factor estimation based on the observation counts of each density in the Gaussian mixture model. In addition, they suggested a simplified non-iterative maximum likelihood scheme for warping factor estimation in training. They found that a low resolution acoustic model (single densities) gave better results than more complex mixture density models. The two-pass approach could be improved by using unnormalized acoustic models for the first recognition pass, which on the other hand increased the gap between two-pass and fast VTN. A word error rate reduction of 9% relative was achieved with Gaussian mixture model based fast VTN on the 5k-word vocabulary Wall Street Journal test set, whereas two-pass VTN yielded up to 17% relative WER reduction. On the German spontaneous speech task VerbMobil I, the reduction was 5% relative at best.

Westphal et al. compared maximum likelihood warping factor estimation with a new criterion based on linear discriminant analysis [Westphal & Schultz$^+$ 98]. The new criterion lead to a faster convergence in iterative warping factor estimation of training data, and the derived speaker cluster were more discriminant. With respect to the word error rate, the new criterion performed slightly worse on the German VerbMobil I task and slightly better on a Chinese dictation task.

Haeb-Umbach investigated in how far cepstral mean normalization and VTN reduce speaker-dependent variations [Haeb-Umbach 99]. He applied Fisher variate analysis to measure inter-speaker variability of phonemes before and after normalization. He found

that not only VTN but also cepstral mean normalization reduces inter-speaker variability by a large extend. In VTN, sentence-wise warping factor estimation compensated for more variations according to this criterion, which was confirmed by lower word error rates in recognition tests on the Wall Street Journal database.

The interaction between VTN and MLLR was further investigated by Uebel and Woodland [Uebel & Woodland 99]. They confirmed that the gain in recognition performance by VTN was smaller than by MLLR, but largely additive in the case of unconstrained MLLR. After several iterations of constrained MLLR, which alone was slightly inferior to unconstrained MLLR, there was no gain observed by VTN. Uebel and Woodland also compared piece-wise linear warping, bi-linear warping, and an approximation to both by linear transformation in the cepstrum domain. They found only little differences in performance between these techniques. Without MLLR, the word error rate could be reduced by up to 6% relative on the SwitchBoard Corpus.

Natio et al. combined phoneme-dependent measures derived from auditory models with speaker-dependent measures derived from a vocal tract model to obtain non-linear warping functions with two free parameters [Naito & Deng$^+$ 99]. Their estimates of the vocal tract length were based on formant frequencies of two specific Japanese vowels. The derived warping functions resembled closely a linear warping functions, which is why only a minor improvement over baseline VTN was achieved on a Japanese phoneme recognition task.

Dolfing evaluated the efficiency of two maximum likelihood criteria for warping factor estimation [Dolfing 00]. One was text-dependent similar to two-pass VTN with the preliminary transcription replaced by the reference transcription in a supervised manner. The other resembled the GMM based text-independent warping factor estimation of Lee and Rose. Based on an internal dictation database he compared the word error rate obtained by these techniques with an optimal error rate by choosing the warping factor with the lowest word error rate. If warping factors were estimated in a speaker-wise fashion, the text-dependent criterion yielded about 90% of the maximum possible reduction in WER. Sentence-wise estimation of the warping factor left more room for improvements, which is why is that case only about 70% of the maximum possible gain was achieved. Preliminary experiments with the text-independent estimation indicated that its performance was only slightly inferior to the text-dependent technique, but conclusive recognition result were not reported.

Cox presented a method to implement VTN at the cepstrum stage [Cox 00]. As Mel-frequency axis warping is approximately a logarithmic scaling of the frequency axis, linear frequency axis warping amounts to a constant frequency shift in the Mel-frequency domain. This fact was used to derive a transformation matrix that compensates for the shift in the cepstrum domain. The functional form of this type of frequency axis warping was similar to highly constrained MLLR with only four free parameters. Phoneme recognition tests in supervised normalization mode using the Wall Street Journal database showed reduced error rates only if the means of single-density acoustic models were adapted. A normalization of the test data did not yield the same improvement. A

minor additional gain was found when a different amount of warping was allowed at different spectral bands. The overall reduction in phoneme error rate was 4% relative at best.

In 2001, Pitz et al. showed that VTN equals a linear transformation in the cepstrum domain for arbitrary invertible frequency axis warping functions [Pitz & Molau$^+$ 01]. This allowed in principle to account for the Jacobian determinant of the transformation, which is typically omitted in maximum likelihood warping factor estimation.

Yet another approach for fast warping factor estimation was presented in the same year by Emori and Shinoda [Emori & Shinoda 01]. They applied bi-linear frequency axis warping in the cepstrum domain and proposed an approximation to compute the warping factor from cepstral coefficients. On a Japanese isolated-word recognition task they achieved similar performance in supervised mode like maximum likelihood estimation at smaller computational costs. Better results were achieved if only vowels were used for estimation, but comparable result for maximum likelihood estimation were not given.

In summary, the following conclusions can be drawn from the previous work:

- There were typically only minor differences between linear and non-linear frequency axis warping functions. For simplicity, linear or piece-wise linear warping was applied in most cases.

- Different implementations of spectral warping were chosen. Similar results were achieved by spectral interpolation or filter bank modification in the frequency domain, or by transformations in the cepstrum domain.

- For warping factor estimation, most groups applied maximum likelihood estimation schemes which had proved to be more robust than estimates based on formant frequencies.

- In training, both iterative and non-iterative methods for warping factor estimation were used.

- In test, two maximum likelihood techniques for warping factors have prevailed: Text-dependent two-pass recognition based on a preliminary transcription from a first recognition pass, and text-independent fast estimation schemes based on simplified acoustic models. The second approach has a much lower computational overhead, but the gain in recognition performance is also lower compared to two-pass VTN.

- With larger training corpora and more complex recognition tasks, the gain achieved by VTN typically decreased from WER reductions well above 10% relative to well below that value.

- The gain in recognition accuracy by VTN and MLLR is to a large extent additive, even though vocal tract length normalization can be viewed as a special case of a highly constrained MLLR. As it relies on only one free parameter, VTN can be applied successfully even if only very little data is available for parameter estimation.

3.1.2 Speaking Rate Normalization

The rate of speech, which is directly linked the process of speech production, is another factor that influences the recognition accuracy of automatic speech recognizers. It was found early that very fast and very slow speakers have on average significantly larger word error rates (e.g. [Pallett & Fiscus$^+$ 94]). There were a number of publications that identified why the recognition accuracy deteriorates for these speakers, how to determine the rate of speech (ROS), and how to improve the recognition of fast speakers. Some publications of general relevance and those related to normalization in the acoustic feature space will be summarized here.

In 1995, Siegler and Stern reported on a number of experiments on the Wall Street Journal corpus aimed at improving the recognition accuracy for fast speakers [Siegler & Stern 95]. They proposed to use the average phoneme rate per second instead of the average word rate. A significantly increased word error rate (more deletions and substitutions) was found for speakers whose rate of speech differed by more than one standard deviation from the mean ROS. The phoneme rate measured on erroneous recognition transcripts was found to be closely linked to the one derived from the reference transcription, though it was systematically lower. Attempts to improve the recognition accuracy by adapting the codebooks to fast speech were unsuccessful. An improved modeling of transition probabilities yielded about 5% relative reduction in word error rate for all speakers. Modifications of the pronunciation dictionary by two simple phonetic rules and by adding compound words again did not improve the recognition accuracy.

Mirghafori et al. presented a detailed analysis why recognition performance degrades for fast speakers [Mirghafori & Fosler$^+$ 95]. Based on the Wall Street Journal and the Resource Management evaluations, they found an increase in word error rate by about a factor of three for the fastest speakers. A mismatch of the acoustic vectors caused by stronger coarticulation in fast speech was verified. Furthermore they found the phonemic duration constraints imposed by their HMM topology to be inappropriate for high rates of speech, where sometimes phonemes are omitted altogether. To cope with the acoustic mismatch, the acoustic model was adapted to the fastest speakers, which yielded a 14% relative reduction in the word error rate on the Wall Street Journal task for fast, but at the same time a 10% relative increase for slow speakers. The temporal constraints of the HMM were relaxed by modified transition probabilities, shorter HMM topologies, and pronunciation variants to model phoneme omissions. Each of these methods gave some improvement for fast speaker, but slightly hurt the overall recognition accuracy. The largest WER reduction for fast speaker of 16% relative was achieved by a combination of emission and transition probability adaptation.

In 1999, Richardson et al. presented cepstrum length normalization as a powerful technique for speaking rate normalization [Richardson & Hwang$^+$ 99]. They defined the rate of speech by the length of an individual phoneme relative to its average duration in the training data, and found that a Gamma distribution described the observed distribution best. To cope with variable rates of speech, they normalized the speaking rate by stretching or squeezing each sentence. Best performance was achieved with

one stretch factor per sentence based on the average phoneme duration calculated in a first recognition pass. According to the stretch factor, new time frames were created in the cepstrum domain by band-limited interpolation between neighboring frames. An alternative modification of the frame shift in signal analysis yielded almost the same result. Experiments on the Wall Street Journal database and an in-house collected data revealed reductions in word error rate of 16% relative for fast speakers and no performance degradation for regular speaking rates. Further experiments showed that the gain in recognition performance achieved by cepstrum length normalization was additive to the gain obtained by unsupervised MLLR.

Pfau et al. investigated different methods to improve the recognition performance on very fast and slow speakers [Pfau & Faltlhauser$^+$ 99]. Speaking rate adapted acoustic models were derived by maximum-a-posteriori (MAP) re-training of the baseline model on speech data of the corresponding category. To reduce variations, maximum likelihood based vocal tract length normalization and pronunciation variants were applied in training and test. In recognition tests on the VerbMobil I corpus with monophone models it was found that pronunciation variants and VTN gave larger that average reductions in word error rate on especially slow and fast speech. A combination of both techniques yielded relative reductions in WER between 16% and 20% in these categories. MAP re-training gave a somewhat smaller gain in recognition accuracy, which was also not additive to the improvements achieved by VTN.

A variable frame rate was proposed by Zhu and Alwan to improve speech recognition in general [Zhu & Alwan 00]. They first derived a large number of acoustic vectors with a small frame shift of 2.5 ms. Next, the Euclidean distance between adjacent frames weighted by the frame energy was computed. Based on the distance, some frames were discarded and others kept. This way formant transitions were described more detailed by more time frames, whereas steady parts of the signal were represented by fewer acoustic vectors. Zhu and Alwan reported improvements on a phoneme recognition task as well as on the TiDigits database. The proposed variable frame rate algorithm proved also to be more robust in conditions with additive noise.

In 2000, Faltlhauser et al. proposed to use Gaussian mixture models for the estimation of the rate of speech, since these had already been successfully applied for gender and speaker recognition as well as in vocal tract length normalization [Faltlhauser & Pfau$^+$ 00]. They trained three GMMs for slow, medium, and fast speech on the VerbMobil I corpus. In two third of all cases the classification of the test data was correct. There were many confusions between neighboring classes but almost none between the two opposite categories. A continuous ROS estimate reasonably close to the phoneme rate determined on the reference transcription was obtained by combining the GMM scores with an artificial neural network.

In the same year, Pfau et al. presented results for a combination of vocal tract length and speaking rate normalization [Pfau & Faltlhauser$^+$ 00]. Variable speaking rate was accounted for by modifying the number of acoustic vectors with linear interpolation in the cepstrum domain. It closely resembled the cepstrum length normalization proposed

by Richardson et al. The gain in recognition performance of both normalization schemes was essentially additive on the VerbMobil I corpus.

In summary, the following conclusions can be drawn from the previous work:

- The number of phonemes or vowels per unit time, and the duration of phonemes relative to their average duration were used as text-dependent measures of the rate of speech.
- Text-independent ROS measures were based on Gaussian mixture models.
- Different methods were proposed to cope with especially fast and slow speakers, some of which were based on normalization in the acoustic feature space.
- The acoustic mismatch could be reduced by adaptation of acoustic models to fast speech, or with speaking rate normalization by cepstrum length interpolation and variable frame shift. In some cases, a performance improvement on fast speech came at the cost of lower recognition accuracy for average or slow speakers.
- The mismatch in HMM duration modeling was handled by adaptation of transition probabilities and shorter HMM topologies.
- Phonetic mismatch was handled by supplementing pronunciation variants with typical phoneme omissions.
- Feature space normalization proved to be successful. The relative reduction in word error rate was typically larger than 10% relative for speakers with especially large deviation from the average rate of speech.
- Vocal tract length normalization was found to be especially successful of very fast and very slow speakers.
- The performance gain by speaking rate normalization was shown to be additive to maximum likelihood linear regression and vocal tract length normalization.

3.1.3 Channel Normalization

The speech waveform produced by a speaker is transmitted over some channel before it reaches the recording device, and the channel disturbs the original speech signal. Convolutional distortions are multiplicative in the spectrum domain. Due to the logarithmic compression before the cosine transform (cf. Section 1.2), multiplicative distortions become additive in the cepstrum domain. Thus, the simplest and most effective case of channel normalization is to subtract the cepstral long-term mean (cepstral mean normalization) which will remove time-invariant distortions introduced by the transmission channel and the recording device. Baseline cepstral mean normalization is nowadays part of virtually every speech recognition system. More advanced channel normalization techniques have been investigated by a number of research groups and shall be introduced in the following.

Huang et al. reported on a number of enhancements of their speech recognition system in 1995 [Huang & Acero$^+$ 95]. One of these was an improved cepstral mean normalization scheme, which was based on independent mean estimates for speech and silence frames. For normalization, the posterior probability of the actual frame to be a speech or silence frame was estimated from the frame energy. Then, the cepstral mean value to be subtracted was derived by linear interpolation between the speech and silence means weighted with the posterior probability. Huang et al. found a minor improvement over baseline cepstral mean subtraction in non-mismatch conditions, but up to 25% relative word error rate reductions in strong mismatch conditions.

Naik pointed out that one of the assumptions for cepstral mean subtraction, namely zero mean for the speech content in the cepstrum, does not hold for short training or test utterances [Naik 95]. Thus, subtracting the cepstral mean will not only remove the channel distortions but also some speech information. Naik proposed a different cepstral mean estimate based on the position of poles in the linear predictive coding (LPC) cepstrum. For two simulated channels it was shown that the pole-filtered channel estimate introduced a smaller error than the cepstral mean estimate. Recognition tests on the Timit database yielded more than 15% reduction in word error rate relative to standard cepstral mean subtraction.

Among other spectral filtering techniques, Junqua et al. investigated the effect of cepstral mean subtraction [Junqua & Fohr$^+$ 95]. They found that subtracting the utterance-wise long-term cepstral mean gives better results than the short-term cepstral mean derived from only a few time frames. Cepstral mean subtraction gave a larger reduction in Mel-frequency cepstrum based signal analysis than relative spectral processing (RASTA). The overall best result on a telephone based name spelling task was achieved with a MFCC front-end including first and second derivatives, and long-term cepstral mean normalization.

In 1997, Westphal studied differed extensions of the standard long-term cepstral mean subtraction [Westphal 97]. He showed that speaker-wise performed better than utterance-wise normalization, and that the efficiency of standard cepstral mean subtraction depends on the silence fraction of the speech signal. To overcome this limitation, it was suggested to compute the cepstral mean on speech frames only, which gave 6% relative WER reduction on the VerbMobil I task, but a slight increase in error rate on the SwitchBoard corpus. Another approach with two separate mean values for speech and silence that were interpolated according to the silence fraction did not perform better than baseline CMS. However, there were some improvements reported on the SwitchBoard task when the difference between the speaker-wise cepstral means for speech and silence and the overall mean obtained on the full database were combined.

In summary, the following conclusions can be drawn from the previous work:

- Even though it is conceptually simple, cepstral mean normalization was found to consistently improve the recognition accuracy in tasks with and without mismatch between training and test.
- Long-term performed better that short-term mean subtraction.
- Improvements over the baseline approach were achieved in some cases when cepstral mean estimates were derived separately on speech and silence frames, and interpolated during normalization.

3.1.4 Noise Suppression

The performance of automatic speech recognition systems usually drops dramatically in the presence of noise. For this reason, noise robustness is a research field of its own which goes beyond the scope of this work. Normalization in the acoustic feature space is one of many possible ways to improve the robustness of speech recognizers [Junqua & Haton 99]. Spectral subtraction [van Compernolle 89], for example, is a standard techniques in noise suppression. It is a model based normalization scheme which relies on the model that the noise is additive in the spectral domain. The contribution of noise is estimated in speech pauses and subtracted from the signal. Some of the feature space matching techniques introduced in Section 3.2.2 can also be regarded as noise suppression schemes.

3.2 Data Distribution Based Normalization Schemes

3.2.1 Feature Space Transformation

Based on statistics of the speech signal, there are some techniques that transform acoustic vectors into a domain that is in general more suitable for automatic speech recognition.

In their 1991 paper, Acero and Stern proposed a number of normalization schemes based on affine cepstrum transformations to counteract a severe degradation of recognition performance when different microphones were used in training and test [Acero & Stern 91]. Some of these techniques required parallel recordings with both microphones in order to estimate the mapping parameters. These techniques were further developed as reported by Liu at al. [Liu & Acero$^+$ 92]. Instead of parallel recordings, a histogram of signal-to-noise ratios was derived for both microphones, and a non-linear transformation function was obtained by histogram mapping. Under strong mismatch conditions, the word error rate could be more than halved.

Neumeyer et al. applied affine transformations of different complexity to the acoustic vector. They compared transformations in the feature and model space, and tested their efficiency on native and non-native speakers on the Wall Street Journal corpus [Neumeyer & Sankar$^+$ 95]. Even though some improvements could be obtained by feature space transformations, they were typically outperformed by adaptation in the model space.

Gopinath showed how to transform multi-dimensional random variables of unknown distribution into Gaussian random variables [Gopinath 00]. The technique called Gaussianization results in uncorrelated and normal distributed dimensions and makes parameter estimation in high-dimensional feature space more robust. Gopinath showed that Gaussianization can be applied as a non-linear feature transformation. Preliminary tests revealed a minor performance improvement on the Broadcast News corpus when Gaussianization was applied instead of logarithmic compression to the filter bank channels.

In summary, affine transformations in the feature space (e.g. at the cepstrum stage) improved the recognition accuracy under mismatch conditions, but were inferior to affine transformations of the acoustic model (adaptation).

3.2.2 Feature Space Matching

Most data distribution based normalization techniques rely on the principal idea of mapping acoustic vectors from the test data space into the training data space to minimize the mismatch between training and test. The techniques proposed in the literature differ mainly in the following ways:

- the domain in which the mismatch is determined (feature or model space)
- the functional form of the transformation (parametric or non-parametric)
- the method for estimating the transformation function (supervised or unsupervised)
- the signal analysis stage at which the feature space is mapped

There have been numerous publications on supervised mapping in the spectral and cepstral domain. In 1992, for example, Matsuko and Hirowo proposed a piecewise-linear mapping between cepstrum vectors from test speakers into a reference space [Matsukoto & Hirowo 92]. The transformation was computed phoneme-wise and later smoothed to maintain continuity in the mapped space. The phoneme error rate of a recognizer with single density monophone models could be reduced significantly by this method.

Neumeyer and Weintraub proposed a piece-wise linear mapping in the cepstrum domain that was unsupervised but relied on a small amount of simultaneous recordings on different channels (e.g. clean and noisy) [Neumeyer & Weintraub 94]. The transformation was based on a set of multi-dimensional linear least-squares filters. Tests were carried out on the Wall Street Journal database transmitted over clean and noisy channels. Whereas under mismatch conditions the word error rate increased by a factor of 2.5 relative to the baseline, it was only a factor of 1.4 after normalization. Neumeyer and Weintraub found almost the same improvements regardless whether clean training data were mapped to the noisy test domain or vice versa.

In 1995, Sankar and Lee investigated transformations in the feature and model space to minimize the mismatch between test utterances and the acoustic model [Sankar & Lee 95]. Their stochastic matching technique was unsupervised and did not require simultaneous recordings. However, it relied on knowledge about the functional

form of the mapping. Sankar and Lee showed that if the transformation was assumed to be a differentiable function with an additional bias term, the EM algorithm could be used to iteratively estimate the parameter of the transformation that maximized the likelihood of the data given the model, and vice versa. Experiments were reported for the Resource Management database with the identity function and an additive bias in the cepstral domain. They showed consistent reductions of the word error rate of well above 50% relative under mismatch conditions (microphone vs. telephone recordings) without cepstral mean normalization. Estimating two bias parameters for speech and silence further improved the recognition performance. In connection with cepstral mean normalization, feature and model space transformations yielded approximately the same reduction in word error rate in the order of 30% relative.

In 1999, Giuliani proposed another unsupervised technique to match the acoustic space of the training and test data, which could be used in online recognition [Giuliani 99]. The main idea was to describe the training and the test data space by two Gaussian mixture models with 128 densities each. If the training data model was used as an initial estimate for the test data model, the update of the densities during re-training could be interpreted as the mismatch between training and test, and used for subsequent feature space matching. The transformation function was an additive term derived by summing up the differences between the training and test GMM densities, weighted by the distance between the density and the current test vector.
Based on Italian speech corpora, Giuliani presented test results for matched and mismatch conditions (close talking vs. hands-free microphone, dictation training data vs. connected digit test data). In the matched case, 17% relative reduction of word error rate was achieved by normalizing the acoustic vectors, which compared to 19% obtained by incremental MLLR. Also in the mismatch case, normalization was slightly inferior to incremental adaptation with word error rate reductions well above 50% relative.

An unsupervised histogram-based mapping technique that also makes no assumption about the functional form of the transformation was proposed in 2000 by Dharanipragada and Padmanabhan [Dharanipragada & Padmanabhan 00]. It was based on the idea of mapping the cumulative distribution of the test data to the cumulative distribution of the training data. Under certain assumptions, this resulted in a simple text-independent histogram matching procedure which was non-parametric, non-linear, and computationally inexpensive. In the case of microphone mismatch, Dharanipragada and Padmanabhan achieved a relative reduction in word error rate of over 30%, which was of the same order as improvements achieved by unsupervised MLLR. The gain of normalization and MLLR was to a large extent additive.

Two years earlier, the same basic technique of matching cumulative distributions was successfully applied in speaker identification by Balchandran [Balchandran & Mammone 98]. Unfortunately, they evaluated the procedure on artificially distorted data only. An additional smoothing factor had to be introduced to avoid over-compensation.

In another publication by Padmanabhan, the histogram mapping technique was further extended [Padmanabhan & Dharanipragada 01]. Linear interpolation between the

points of the non-linear mapping function reduced quantization errors. In addition, a text-dependent extension was proposed, in which the mapping function was estimated in a maximum likelihood framework. The aim was to get robust estimates of the transformation with only a few adaptation sentences. On the same test corpus as in their previous paper, a minor improvement over supervised MLLR was achieved with the maximum likelihood based transformation function.

Hilger and Ney applied a parametric histogram normalization technique at the filter bank stage [Hilger & Ney 01]. Only four bins (quantiles) of the cumulative histograms were estimated, and piece-wise linear and power transformation functions were fitted to these bins according to a minimum squared error criterion. Normalization was applied in test only, and the reference histogram was averaged over all filter channels on the training data. It was shown that even single word utterances were sufficient to estimate the transformation function reliably, which made this technique useful for online applications. The power function turned out to be more robust than the piece-wise linear transformation. Recognition tests on a number of noisy speech corpora (recordings in car environment) yielded significant performance improvements proportional to the degree of mismatch between training and test. Histogram normalization proved to be superior to other noise suppression schemes investigated by the authors (e.g spectral subtraction).

In summary, the following conclusions can be drawn from the previous work:

- Many feature space matching procedures proposed in the literature were either supervised or required simultaneous recordings from the different environments. Under these idealized conditions, a large gain in recognition performance could be obtained.
- The transformation parameters were estimated in different spaces: Most supervised techniques as well as the histogram-based methods relied on distributions of the training and test data. The stochastic matching of Sankar and Lee transforms the test data to better match the acoustic model, and the Gaussian mixture model based approach of Giuliani derives the transformation function from simplified acoustic models for training and test data.
- The stochastic matching technique of Sankar and Lee is unsupervised but makes assumptions about the functional form of the transformation.
- The Gaussian mixture model based approach of Giuliani relies on the assumption that the mismatch can be expressed by deviations of prototype vectors describing the training and test data space.
- The histogram-based method of Dharanipragada and Padmanabhan is unsupervised and non-parametric. As it relies on global statistics of the speech data, a larger amount of adaptation data is required to estimate the transformation reliably.
- If the number of histogram bins is reduced and a parametric transformation function is fitted to the discrete histogram points, histogram normalization can be applied successfully even if only little adaptation data (e.g. single words) are available.
- Performance gain by feature space matching is to some extent additive to adaptation schemes like maximum likelihood linear regression.

Chapter 4

Aims of this Work

Based on the RWTH large vocabulary speech recognition system (described in [Ney & Welling$^+$ 98] and [Sixtus & Molau$^+$ 00], for example), state-of-the-art model based and data distribution based normalization techniques in the acoustic feature space shall be developed, studied, and improved.

Long-term *cepstral mean normalization* has proved to be an efficient scheme for channel normalization (cf. Section 3.1.3). It will be a fixed part of the baseline system. Normalization takes place on a sentence-wise level unless there are demands for online recognition as in the VerbMobil II task, where a sliding window will be used.

In addition, cepstral variance normalization will be part of the baseline system for those tasks were it helps to reduce the word error rate.

Speaking rate normalization will not be further pursued in this work. These techniques typically improve the recognition performance for fast or slow speakers, but do not help for the majority of "average" speakers (cf. Section 3.1.2). In addition, it was shown that vocal tract length normalization, which will be one of the focal points in this work, is especially effective on critical speakers with a much higher or lower than average rate of speech.
In the case of conversational speech as in the VerbMobil II corpus, the HMM topology of the baseline system will be modified for improved recognition of fast speech (cf. Section 1.3).

Vocal tract length normalization in the maximum likelihood framework is a well-established normalization scheme that proved to be more or less effective on every database it was applied to (cf. Section 3.1.1). Some of the results of other groups shall be validated and a number of open issues shall be addressed in Chapter 6 this work:

I Whereas the reduction of word error rate is typically well above 10% relative in the case of "simple" tasks or acoustic models (acoustic modeling that is not state-of-the-art, training on small databases, small vocabulary), it reduces to well below 10% for large vocabulary systems with advanced acoustic modeling trained on a large amount of data. The baseline text-dependent two-pass VTN scheme shall be optimized in such a way that it consistently yields a significant reduction in word

error rate in the order of 10% relative on difficult tasks and in variable environments. Among the issues to be addressed are improved estimates of the warping factor on speech frames, the influence of the warping function, re-estimation of the phonetic decision tree and the LDA transformation matrix on normalized data, and iterative warping factor estimation.

II Different text-independent approaches for fast warping factor estimation were proposed in the literature. These proved to approximately halve the computation time in comparison to two-pass VTN, but in most cases the gain in recognition performance reduced significantly at the same time. A comparison of different Gaussian mixture model based warping factor estimation schemes shall be carried out. The aim is to find a method that gives the same optimum recognition performance as two-pass VTN without an increase in the real-time factor compared to the baseline system without vocal tract length normalization.

III For application in online systems it is not only desirable to have a low word error rate, but also only little or no delay between speech recording and the recognition output. Text-independent warping factor estimation shall be modified to minimize the time delay introduced by signal analysis, and it shall be proven that VTN can be applied successfully in online speech recognition.

IV Different ways of implementing frequency axis warping in either the time, the frequency, or the cepstrum domain were proposed in the literature. Many of these techniques are either limited to certain frequency axis warping functions, or they possibly suffer from problems like quantization and interpolation errors, or a modified bandwidth. A simplified signal analysis front-end shall be proposed that integrates all spectral warping (Mel-frequency and VTN warping) into the discrete cosine transform.

Among the data distribution based normalization schemes, *histogram normalization* seems to be the most promising one (cf. Section 3.2.2). Histogram normalization is widely used in image processing, but the application in automatic speech recognition is still largely unexplored. It is an unsupervised technique that does not require simultaneous recordings and makes no assumption about the functional form of the transformation. It was shown that histogram normalization significantly improves the recognition accuracy in mismatch conditions, and that it is computationally attractive. There are a number of open issues and limitations that shall be addressed in Chapter 7 of this work:

V So far, histogram normalization has been applied at either the filter bank stage or to the final acoustic vector (in this case at the cepstrum stage). It has not been investigated before, at which stage in signal analysis histogram normalization is most appropriate. Hence, this technique shall be applied at all possible stages to find out where it is most effective, and sequential normalization at different stages shall be investigated as well.

VI All results published so far were achieved with only the test data being normalized. As a normalization technique, however, additional gain in recognition accuracy is to be expected, when both training and test data are mapped to the same reference

condition (cf. Section 2.2). The effect of histogram normalization in training and test shall be investigated in this work.

VII Histogram normalization is based on two assumptions about the global statistics of the speech signal and the orientation of the feature space axes. The first assumption shall be relaxed by considering the variable silence fraction in the utterances. Furthermore, a new rotation based normalization scheme shall be introduced that matches the principal feature space axes with the largest data scatter and overcomes the second assumption of histogram normalization.

VIII In previous work, feature space matching has proved to be an effective way of coping with large mismatch between training and test data (e.g. different microphones, recordings in clean vs. noisy environments). However, it might also reduce speaker and channel dependent variations in the speech signal. For this purpose, histogram normalization and feature space rotation shall be evaluated on corpora with different degrees of mismatch.

IX It is to expect that the gain in recognition performance obtained by vocal tract length normalization and histogram normalization/feature space rotation is to some extent additive as both account for different variations (frequency shifts vs. different global data distribution) and have a different functional form. To which extent the gain is additive shall be examined in this work.

Development will take place in the framework of a within-word system, as at the time of evaluation of most algorithms an across-word system was under development and not yet established at RWTH. In the end, the proposed techniques will also be tested with a system including across-word triphone models [Sixtus 02].

Chapter 5

Corpora and Recognition Setup

5.1 Introduction

The normalization methods studied in this work will be evaluated on a variety of corpora with different vocabulary sizes, complexities (isolated words vs. continuous speech), acoustic conditions (office vs. telephone and car recordings), and speaking styles (planned vs. spontaneous speech). The corpora will be introduced in detail in the following sections together with the baseline recognition setup optimized for each task.

5.2 Clean Acoustic Conditions

5.2.1 North American Business News

The *North American Business News* (*NAB*) corpus consists of business texts from the Wall Street Journal (*WSJ*). The texts were read by journalists and recorded under clean studio conditions [Pallett & Fiscus$^+$ 93]. The WSJ0 and WSJ1 training corpora were collected by the American *National Institute of Standards and Technology*.

Recognition tests will be carried out on the NAB November '94 H1 development test set with a 20k-word vocabulary [Pallett & Fiscus$^+$ 95][Kubala 95]. The test set contains 2.7% unknown words. The statistics of the training and test corpora are summarized in Table 5.1. Average condition duration is the amount of data available for histogram and covariance matrix estimation discussed in Chapter 7. The recognition setup can be summarized as follows:

- 20 filter bank channels
- 16 Mel-frequency warped cepstral coefficients
- sentence-wise long-term cepstral mean normalization and energy normalization
- LDA on seven adjacent MFCC vectors, reduction to 32 dimensions
- 3000/7000 decision-tree based generalized within-word/across-word triphone states plus one silence state

Table 5.1: Statistics of the NAB 20k training and test corpora.

Corpus	Training WSJ0+1	Test DEV-94 H1
Language	English	
Speaking Style	planned	
Bandwidth	microphone	
Overall Duration [h]	81.4	0.8
Silence Fraction [%]	27	19
Average Condition Duration [s]	796	146
# Speakers	284	20
# Sentences	37 571	310
# Running Words	643 754	7 387
# Running Phonemes	2 685 482	-
Trigram LM Perplexity	-	124.5

- gender-independent acoustic within-word/across-word models with up to 597k/694k Gaussian mixture densities
- 6-state HMM topology
- trigram language model

5.2.2 VerbMobil II

VerbMobil was a German speech-to-speech translation research project (phase I: 1993-96, phase II: 1997-2000) for spontaneous speech in the domain appointment scheduling (scenario A) and information desk (scenario B) [Wahlster 00]. The joint effort of numerous universities, speech technology companies, and research institutes was funded by the German *Ministry for Education, Science, Research and Technology* (BMBF). VerbMobil covered three languages, namely German, English and Japanese.

In the second project phase, VerbMobil was extended by a remote PC maintenance task (scenario C) to allow comparison with existing commercial speech technology products. At that time, the *Lehrstuhl für Informatik VI* of *RWTH Aachen* contributed a German LVCSR system [Sixtus & Molau$^+$ 00][Kanthak & Sixtus$^+$ 00] that will be used for the experiments reported in this work. In addition, a speaker-dependent recognizer for scenario C and a probabilistic machine translation module were provided by the *Lehrstuhl für Informatik VI.*

In the framework of the VerbMobil project, a corpus of German spontaneous speech was collected and annotated [Burger & Weilhammer$^+$ 00]. Speech data were gathered at different sites over three channel types:

- close-talking microphones
- room microphones
- various telephone channels (mobile, wireline, wireless)

For the experiments reported here, German microphone training data from all three scenarios will be used. The corpus consists of 49 hours of speech data collected with a close-talking microphone, and 14 hours collected with a room microphone.

Recognition tests will be carried out with a 10k-word vocabulary on two different speaker-independent corpora, the 1999 development test corpus for scenario A and B (DEV99AB) with an out-of-vocabulary (OOV) rate of 1.9%, and for scenario B alone (DEV99B, OOV rate = 2.0%). Both were recorded with close-talking microphones.

The second corpus is a subset of the first. It is characterized by a domain mismatch (the majority of training data were collected in scenario A) and a minor acoustic mismatch (most training data were collected with a different close-talking microphone type). The statistics of the training and test corpora are summarized in Table 5.2.

For the rapid development and test of algorithms as well as for experiments under mismatch conditions, the smaller development corpus DEV99B will be employed. Final results for speaker normalization will be given for the full DEV99AB corpus.

The VerbMobil II recognition setup can be summarized as follows:

- 20 filter bank channels

Table 5.2: Statistics of the VerbMobil II training and test corpora.

Corpus	Training	Test	
	CD1-41	DEV99B	DEV99AB
Language	German		
Speaking Style	spontaneous		
Bandwidth	microphone		
Overall Duration [h]	61.5	0.5	1.6
Silence Fraction [%]	13	11	11
Average Condition Duration [s]	140	112	112
# Speakers	857	6	16
# Sentences	36 010	336	1 081
# Running Words	560 837	4 346	14 662
# Running Phonemes	2 308 741	18 040	59 820
Class-Trigram LM Perplexity	-	74.6	62.0

- 16 Mel-frequency warped cepstral coefficients, 16 first derivatives, second derivative of the energy
- short-term cepstral mean and variance normalization in a symmetric sliding window of two seconds length
- LDA on three adjacent MFCC vectors including the derivatives, reduction to 33 dimensions
- 2500/3500 decision-tree based generalized within-word/across-word triphone states plus one silence state
- gender-independent within-word/across-word acoustic models with up to 456k/579k Gaussian mixture densities
- 3-state HMM topology
- class-trigram language model as of October 1999

Informal tests have shown that the class trigram language model used for the final VerbMobil II evaluation yields consistent reductions in word error rates of about 1.5% absolute on the DEV99AB corpus, and 2.8% absolute on the DEV99B corpus compared to the language model used here. However, this has no impact on the evaluation of the algorithms that will be presented in this work.

5.3 Degraded Acoustic Conditions

5.3.1 EuTrans II

EuTrans was a research project on example-based machine translation techniques for text and speech input in a traveler task domain (phase I: 1996, phase II: 1997-2000) [Casacuberta & Llorens$^+$ 01]. The project was supported by the European Union ESPRIT Long Term Research Program with an overall of four project partners located in Spain, Italy, and Germany. EuTrans covered three languages, namely Italian, Spanish, and English.

One work package involved the collection of an Italian spontaneous speech corpus over wireline telephone [di Carlo 00]. In the second phase of the project, the *Lehrstuhl für Informatik VI* of *RWTH Aachen* contributed acoustic models trained on this corpus, and a probabilistic machine translation system.

The telephone training and test data were collected in the same environment, but the channel quality varied significantly between different recording sessions. Recognition tests will be carried out with a 2k-word closed vocabulary on the final evaluation test set. The statistics of the training and test corpora are summarized in Table 5.3. The EuTrans II recognition setup can be summarized as follows:

Table 5.3: Statistics of the EuTrans II training and test corpora.

Corpus	Training D1.3c/d	Test EVAL00
Language	Italian	
Speaking Style	spontaneous	
Bandwidth	telephone	
Overall Duration [h]	7.9	0.8
Silence Fraction [%]	32	33
Average Condition Duration [s]	104	119
# Speakers	276	25
# Sentences	3 187	300
# Running Words	52 700	5 555
# Running Phonemes	250 749	26 853
Trigram LM Perplexity	-	28.6

- 15 filter bank channels
- 12 Mel-frequency warped cepstral coefficients, 12 first derivatives, second derivative of the energy
- sentence-wise long-term cepstral mean normalization, variance normalization, and energy normalization
- LDA on three adjacent MFCC vectors including the derivatives, reduction to 25 dimensions
- 1500/3000 decision-tree based generalized within-word/across-word triphone states plus one silence state
- gender-independent within-word/across-word acoustic models with up to 96k/121k Gaussian mixture densities
- 6-state HMM topology
- trigram language model

5.3.2 CarNavigation

CarNavigation is a German isolated-word database that was collected by the *Lehrstuhl für Informatik VI* of *RWTH Aachen* for *Panasonic Technology Inc. / Speech Technology Laboratory.*

The training data were recorded in a quiet office environment with a close-talking microphone, and they consist of isolated words and spelling sequences.

The closed-vocabulary test sets consist of isolated-word utterances recorded in various environments. The office test set was collected under the same conditions as the training

data. Two other test sets were recorded in cars (city and highway traffic) with the speaker sitting on the passenger seat and the microphone mounted above the speaker on the visor.

Each test set consists of 2 100 equally probable unique words which were uttered only once. There is no overlap in vocabulary among the test sets, and between the training and test corpora. Because of bad recording quality, 31 utterances were removed from the office test set, but the 2 100 word recognition vocabulary remained the same.

The statistics of the training and test corpora are summarized in Table 5.4. The CarNavigation recognition setup can be summarized as follows:

- 20 filter bank channels
- 16 Mel-frequency warped cepstral coefficients
- short-term cepstral mean and (optional) variance normalization in a symmetric sliding window of two second length
- LDA on nine adjacent MFCC vectors, reduction to 33 dimensions
- 700 decision-tree based triphone states plus one silence state
- gender-independent acoustic models with up to 22k Gaussian mixture densities
- 6-state HMM topology
- zerogram language model

Pruning (cf. Section 1.5) will be deactivated (full search) and the recognizer will be forced to recognize exactly one word for each test utterance.

Table 5.4: Statistics of the CarNavigation training and test corpora.

Corpus	Training	Test		
	Office	Office	City	Highway
Language	German			
Speaking Style	planned			
Bandwidth	microphone			
Overall Duration [h]	18.8	1.7	1.7	1.8
Silence Fraction [%]	60	69	73	75
Average Condition Duration [s]	785	425	450	468
Average SNR [dB]	21	21	9	6
# Speakers	86	14	14	14
# Running Words	61 742	2 069	2 100	2 100
# Running Phonemes	189 996	16 842	17 184	17 117
Vocabulary	-	2 100	2 100	2 100

Chapter 6

Vocal Tract Length Normalization

6.1 Motivation

Vocal tract length normalization is a model based normalization scheme (cf. Section 2.4) that relies on a model of speech production by the human speech apparatus, in particular on the size of the vocal tract (Figure 6.1).

The vocal tract, i.e. the position and shape of the different organs of speech, determines the sound that is generated. The simplest model for the vocal tract is a uniform tube of length l between the vocal cords and the lips [Eide & Gish 96]. The tube is closed at one and open at the other end (Figure 6.2).

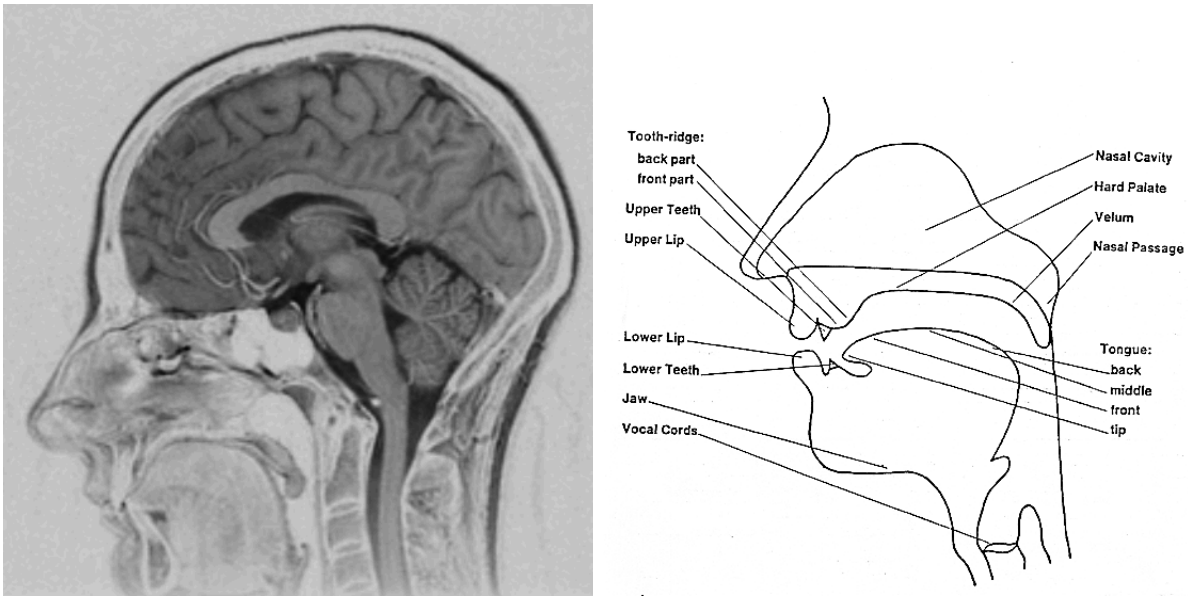

Figure 6.1: Photograph of a mid-sagittal section of a human head (left) and a schematic plot of the organs of speech (right). The pictures were taken from http://www.phon.ox.ac.uk/∼jcoleman/phonation.htm

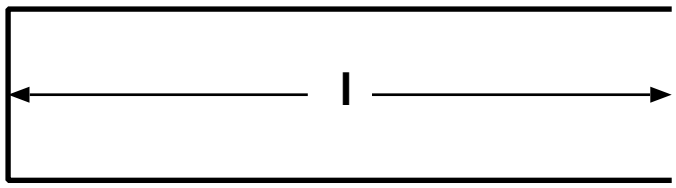

Figure 6.2: The human vocal tract can be modeled by a uniform tube of length l that is open at one end.

According to this model, formant center frequencies of the speech signal are inverse proportional to the length of the vocal tract as they occur at odd multiples of l^{-1}. Since the vocal tract length varies from about 13 cm for female to over 18 cm for male speakers, there are systematic inter-speaker variations of formant frequencies by up to 25% [Lee & Rose 96]. These variations are irrelevant and to some extend harmful to automatic speech recognition.

Gender-dependent modeling is one way of coping with variable vocal tract lengths. By training one acoustic model for female speakers with short vocal tracts and another for male speakers with long vocal tracts, most of the formant frequency variations are accounted for. Thus, gender-dependent acoustic models cover more relevant information of the speech signal which is why they are superior to gender-independent models. However, as discussed in Section 2.1, splitting of training data among different acoustic models has certain disadvantages compared to transformations.

The idea of vocal tract length normalization is to transform the speech signal by some function $f_\alpha(\cdot)$ such that the mean formant frequencies for each training and test condition match those of the reference condition (cf. Section 2.3). In this case, conditions are synonymic to speakers, since the length of the vocal tract is speaker-dependent. The reference condition is defined as the mean over all training speakers. Hence, the aim of VTN is to match the mean formant frequencies of each speaker to the average formant frequencies over all training speakers.

According to the tube model, formant frequencies are shifted downward linearly with increasing vocal tract lengths. A straight-forward solution to account for that shift is to warp the frequency axis during signal analysis (Figure 6.3). Other approaches working in the time or cepstral domain can achieve a similar effect (cf. Section 3.1.1). In most cases, the transformation function $f_\alpha(\cdot)$ depends on a single parameter α only, which is called warping factor and describes the amount of spectral warping required for each speaker.

The basic idea of vocal tract length normalization is simple, but its efficient implementation is not. Most challenging is to estimate the warping factor from data. On the one hand, the algorithm needs to be robust and give reliable estimates from only little speech data. On the other hand it should not introduce too much computational overhead to the speech recognition system.

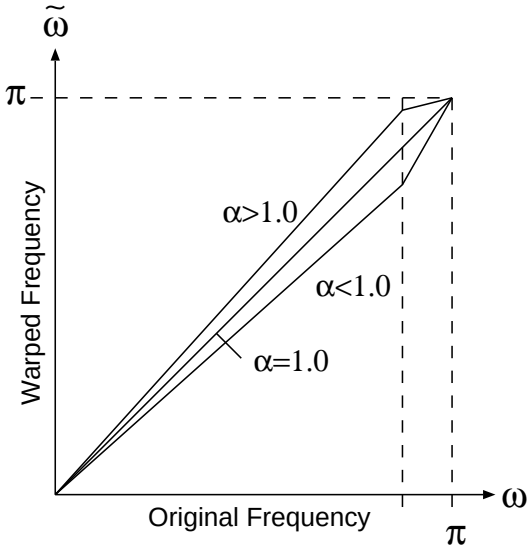

Figure 6.3: Principle of vocal tract length normalization: The frequency axis of the speech signal is warped during signal analysis. Here, a piece-wise linear warping function with warping factor α is depicted.

Maximum likelihood based techniques have prevailed for warping factor estimation (cf. Section 3.1.1). In theory, it would be preferable if parameter estimation in training and test were identical, but in practice the situation in training is different from test. In training, the reference transcription is given, but no normalized acoustic model. In test, the transcription in unknown, but the warping factors from the training speakers and the corresponding normalized acoustic model are available. Hence, different parameter estimation techniques are usually applied.

The other aspect of vocal tract length normalization is to find a suitable method for spectral warping, and to incorporate the transformation efficiently into the architecture of a speech recognition system.

In the next section, the baseline procedure of warping factor estimation in training will be described. Then a number of alternative estimation schemes in test will be studied. Evaluation criteria are the word error rate and the real-time factor. A number of optimizations will be proposed to maximize the gain in recognition performance, and an optimal training procedure based on the previous experiments will be proposed as conclusion.

6.2 Warping Factor Estimation in Training

Like most normalization techniques, vocal tract length normalization gives best performance when both test and training data are normalized (cf. Section 2.2). According to

Eqn. 2.8, the normalized acoustic model $\tilde{\theta}$ should be derived by a joint optimization over the unknown warping factors and the unknown acoustic model parameters. Assuming a uniform prior distribution and neglecting the Jacobian determinant, this leads to the following training criterion:

$$\begin{aligned}\tilde{\theta} &\cong \arg\max_{\theta} \prod_{r=1}^{R} \max_{\alpha} \{p(X_r|W_r;\theta,\alpha)\} \\ &\cong \arg\max_{\theta} \prod_{r=1}^{R} \max_{\alpha} \{\tilde{p}(X_r^{\alpha}|W_r;\theta)\} \end{aligned} \tag{6.1}$$

Here X_r^{α} denotes the acoustic vectors from speaker r normalized with warping factor α.

For the joint optimization, an iterative training procedure has been proposed in the literature [Lee & Rose 96]. In each iteration, first the acoustic model parameters are estimated by maximum likelihood with fixed warping factors from the previous iteration. Then, the warping factors are re-estimated with the updated acoustic model. By splitting the training database it is ensured that each warping factor α_r is estimated with an acoustic model that was trained on data excluding speaker r.

Alternatively, the warping factor $\hat{\alpha}_r$ for each training speaker r may be estimated beforehand by some function $h(\cdot)$ (cf. Section 2.2 and Eqn. 2.9) and kept fixed during acoustic model training (Eqn. 2.10) to avoid the complex joint optimization:

$$\begin{aligned}\hat{\alpha}_r &= h(X_r, W_r, \theta_0) \\ &= \arg\max_{\alpha} p(X_r|W_r;\theta_0,\alpha) \\ &\cong \arg\max_{\alpha} \tilde{p}(X_r^{\alpha}|W_r;\theta_0) \end{aligned} \tag{6.2}$$

The conditional probability of Eqn. 6.2 is computed by forced alignment. A grid search is carried out for different discrete values of the warping factor α, because a closed-form solution of the optimization criterion is not obvious [Lee & Rose 96].

Since the reference transcription is known in training, the only free parameter is the initial acoustic model θ_0 used for warping factor estimation. In principle it should be a normalized acoustic model to match with the normalized acoustic vectors, but as long as there are no warping factors for training speakers there is also no normalized acoustic model. One solution is the iterative training procedure described above, but in practice it turned out that more than one training iteration was not helpful, anyway [Lee & Rose 96].

Welling proposed to use an unnormalized acoustic model of low resolution [Welling & Kanthak$^+$ 99]. He found that a single density model is the best choice for θ_0, since mixture density models have already adapted to different warping factors and do not discriminate well between them anymore. A low resolution model captures only general properties of the speech signal. It nicely approximates the reference condition

(defined as the average over all training speakers, cf. Section 6.1) and resembles closely a normalized acoustic model. θ_0 is obtained by conventional maximum likelihood training on unnormalized data:

$$\theta_0 = \arg\max_{\theta} \prod_{r=1}^{R} p(X_r|W_r;\theta) \tag{6.3}$$

Once the warping factors are estimated, the training data are normalized (Eqn. 6.4) and kept fixed during the training of the normalized acoustic model $\tilde{\theta}$ (Eqn. 6.5):

$$X_r \rightarrow \tilde{X}_r = f_\alpha(X_r) = X_r^{\hat{\alpha}_r} \tag{6.4}$$

$$\begin{aligned} \tilde{\theta} &\cong \arg\max_{\theta} \prod_{r=1}^{R} p(X_r|W_r;\theta,\hat{\alpha}_r) \\ &\cong \arg\max_{\theta} \prod_{r=1}^{R} \tilde{p}(X_r^{\hat{\alpha}_r}|W_r;\theta) \end{aligned} \tag{6.5}$$

In this work, the training procedure proposed by Welling is adopted. Iterative re-estimation of warping factors will be examined in Section 6.4.4.

6.3 Warping Factor Estimation in Test

The situation in test is complicated as well, because according to Bayes' decision rule for adaptive acoustic modeling (Eqn. 2.7) there should be a joint optimization over the unknown word sequence W and the unknown warping factor α. Assuming the prior distribution $p(\alpha|W;\theta)$ to be uniform and neglecting the Jacobian determinant yields:

$$\begin{aligned} W &\cong \arg\max_{W'} \left\{ p(W') \cdot \max_{\alpha} p(X|W';\tilde{\theta},\alpha) \right\} \\ &\cong \arg\max_{W'} \left\{ p(W') \cdot \max_{\alpha} \tilde{p}(X^{\alpha}|W';\tilde{\theta}) \right\} \end{aligned} \tag{6.6}$$

Contrary to training, the warping factors of the training speakers and the normalized acoustic model $\tilde{\theta}$ are given, but the word sequence W is unknown. Furthermore, the identity of the test speaker is usually unknown, which is why warping factors in test have to be estimated sentence-wise.

In the following sections a number of solutions of different complexity will be studied. The computational overhead will be investigated in terms of the size of the search space (average number of active states after histogram pruning) and the real-time factor (RTF) with fixed pruning thresholds, measured on a 600 MHz Pentium III PC with 1 GB main memory.

6.3.1 Full Optimization

If computation time is not an issue, one solution is to do a full optimization of the recognized word sequence W and the warping factor α (cf. Section 2.1):

$$\max_{\alpha}\{\max_{W}\{p(W)\cdot\tilde{p}(X^{\alpha}|W,\tilde{\theta})\}\} \tag{6.7}$$

In this case, an own recognition pass has to be carried out for each considered warping factor, and the word sequence that corresponds to the maximizing warping factor is selected. If n different warping factors are considered, the real-time factor will increase by a factor of the order of n.

Recognition results for full optimization are summarized in Table 6.1 in the following section.

6.3.2 Text-Dependent Warping Factor Estimation

To avoid an expensive full optimization, the warping factor α may be determined beforehand by some function $h(\cdot)$ and kept fixed in the following optimization over the unknown word sequence W.

In the baseline two-pass recognition approach, $h(\cdot)$ is chosen text-dependent similar to training (cf. Eqn. 6.2). Since the spoken word sequence W is unknown in test, it is replaced by the preliminary word sequence $\hat{W}$ determined in a first recognition pass on unnormalized data (Eqn. 6.8). Furthermore, the low resolution acoustic model θ_0 is replaced by the normalized mixture-density model $\tilde{\theta}$. Poor discrimination between warping factors (cf. Section 6.2) is no longer a problem, as the model was trained on normalized data.

Two-pass recognition can be summarized as follows:

1. Determine a preliminary transcription $\hat{W}$ in a (non-adaptive) first recognition pass:

$$\hat{W} = \arg\max_{W}\{p(W)\cdot p(X|W,\theta)\} \tag{6.8}$$

2. Similar to training, perform a forced alignment for each considered warping factor $\hat{\alpha}$. Choose the one that maximizes the conditional probability $\tilde{p}(\cdot)$ given the preliminary transcription and the normalized acoustic model:

$$\begin{aligned}\hat{\alpha} &= h(X,\hat{W},\tilde{\theta}) \\ &= \arg\max_{\alpha}\tilde{p}(X^{\alpha}|\hat{W};\tilde{\theta})\end{aligned} \tag{6.9}$$

3. Second (adaptive) recognition pass on normalized acoustic vectors:

$$W = \arg\max_{W'}\{p(W')\cdot\tilde{p}(X^{\hat{\alpha}}|W';\tilde{\theta})\} \tag{6.10}$$

Welling found that best performance is achieved when an unnormalized acoustic model θ is used in the first recognition pass [Welling & Kanthak$^+$ 99]. Given the concept of adaptation and normalization developed in Section 2.2 this comes as no surprise, as a model trained on different conditions performs better under the variety of test conditions than a normalized acoustic model.

For test purposes, the preliminary transcription $\hat{W}$ in Eqn. 6.9 may be replaced by the reference transcription in a supervised manner. This technique is not applicable in real applications, but it allows to study the dependence of the recognition performance on errors in the preliminary transcription.

Recognition test results on the VerbMobil II corpus for full optimization, two-pass VTN, and supervised VTN are summarized in Table 6.1. As expected, supervised VTN yields best results with respect to the recognition accuracy.

Two-pass VTN gives about the same word error rate as the full optimization, and both are only little inferior to supervised vocal tract length normalization. Hence, the warping factor can be reliably estimated beforehand even if the word error rate in the first recognition pass is of the order of 25%. The reason is that contrary to most adaptation techniques just one parameter has to be estimated in vocal tract length normalization, which requires very few data. In addition, mis-recognized words are often phonetically similar to the correct words, hence, the phoneme error rate is significantly lower than the word error rate [Wessel & Ney 01].

An improved handling of silence frames introduced in Section 6.4.1 will level out the remaining difference between two-pass and supervised warping factor estimation.

Two-pass vocal tract length normalization requires approximately twice as much computation time as conventional non-adaptive recognition due to the two recognition passes. For the full optimization, 13 recognition passes were required for the different warping factors. The real-time factor increased by a factor of 14, because in most cases recognition took place with an inadequate warping factor which lead to a strong mismatch and an increase of the search space by 7% on average.

Table 6.1: Recognition test results on the VerbMobil II DEV99B corpus for different VTN schemes in test. Given are the average number of active states after histogram pruning, the real-time factor, and the word error rate.

VTN	Search Space		Overall [%]	
Warping Factor Estimation	States	RTF	Del - Ins	WER
baseline without normalization	5792	13.4	4.1 - 4.7	24.9
full optimization	6214	178.2	4.3 - 4.8	23.0
two-pass	5050	26.2	4.1 - 4.7	22.8
supervised	5032	12.5	4.1 - 4.5	**22.3**
cheating	6214	178.2	3.7 - 3.2	18.4

The last line in Table 6.1 gives the result of a cheating experiment. Based on the full optimization, the warping factor was selected for each sentence that minimized the word error rate. The experiment shows that if the best warping factor could be guessed somehow, the word error rate would drop by more than 25% relative. However, this result should not be compared with the other warping factor estimation techniques, but with n-best-list or word graph error rates: For each sentence, the best of up to 13 alternative word sequences is chosen.

With two-pass or even supervised VTN, 40% or less of the maximum possible gain in recognition performance is obtained as indicated by the cheating experiment. This contradicts the result of Dolfing, who reported up to 70% of the possible improvements for supervised vocal tract length normalization with sentence-wise warping factor estimation [Dolfing 00]. A closer examination reveals that his frequency axis warping scheme is presumably sub-optimal. The improvements by supervised VTN relative to the baseline are comparable, but the cheating experiment of Dolfing reduced the word error rate by a much smaller amount.

6.3.3 Text-Independent Warping Factor Estimation

The two-pass VTN approach performs as well as the full optimization but still doubles the real-time factor as it requires a preliminary transcription from a first recognition pass. This makes it difficult to apply two-pass vocal tract length normalization in online applications.

Text-independent techniques that will be presented in the following are based on the observation that the global distribution of acoustic vectors in the feature space varies for speakers with different warping factors independently of what is spoken. Hence, by modeling the feature space with simplified acoustic models it will be possible to use a text-independent function $h(\cdot)$ for warping factor estimation (cf. Eqn. 2.9), and save the first recognition pass. These techniques, which are otherwise similar to the text-dependent warping factor estimation, are summarized as fast vocal tract length normalization.

Wegmann et al. and Welling et al. suggested to model the distribution of normalized acoustic vectors $\tilde{X}$ by a Gaussian mixture model (GMM) $\tilde{\Lambda}$ [Wegmann & McAllaster$^+$ 96] [Welling & Haeb-Umbach$^+$ 98]. After the warping factors $\hat{\alpha}_r$ are determined for each training speaker r as described in Section 6.2, the Gaussian mixture model is trained by maximum likelihood:

$$\begin{aligned} \tilde{\Lambda} &\cong \arg\max_{\Lambda} \prod_{r=1}^{R} p(X_r|\Lambda, \hat{\alpha}_r) \\ &\cong \arg\max_{\Lambda} \prod_{r=1}^{R} \tilde{p}(X_r^{\hat{\alpha}_r}|\Lambda) \end{aligned} \tag{6.11}$$

The covariance matrix Σ of the Gaussian mixture model $\tilde{\Lambda}$ is diagonal and pooled over all L densities (Eqn. 6.12). The densities are weighted by normalized mixture weights $c_1, \ldots, c_L$. As usual, the maximum approximation is applied at the density level for faster likelihood calculations (Eqn. 6.13):

$$\tilde{p}(x|\tilde{\Lambda}) = \sum_{l=1}^{L} c_l \cdot \mathcal{N}(x|\mu_l, \Sigma; \tilde{\Lambda}) \qquad \sum_{l=1}^{L} c_l = 1 \tag{6.12}$$

$$\cong \max_l \{c_l \cdot \mathcal{N}(x|\mu_l, \Sigma; \tilde{\Lambda})\} \tag{6.13}$$

In test, the acoustic vectors are normalized with different warping factors, and the one with maximum likelihood is selected for recognition:

1. Find the warping factor $\hat{\alpha}$ that maximizes the likelihood given the simplified acoustic model $\tilde{\Lambda}$:

$$\begin{aligned} \hat{\alpha} &= h(X, \tilde{\Lambda}) \\ &\cong \arg\max_{\alpha} \tilde{p}(X^{\alpha}|\tilde{\Lambda}) \end{aligned} \tag{6.14}$$

2. Adapted recognition pass as in two-pass VTN (cf. Eqn. 6.10)

And alternative approach was suggested by Lee and Rose [Lee & Rose 96]. It is based on a set of Gaussian mixture models Λ_α. Each of them describes the distribution of unnormalized acoustic vectors X_r from all speakers r with a specific warping factor $\alpha_r = \alpha$:

$$\Lambda_\alpha = \arg\max_{\Lambda} \prod_{r:\alpha_r=\alpha} p(X_r|\Lambda) \tag{6.15}$$

In test, the likelihood of the acoustic vector sequence is calculated with all Gaussian mixture models, and the warping factor that corresponds to the model with maximum likelihood is selected for recognition:

1. Find the warping factor $\hat{\alpha}$ that maximizes the likelihood given the set of Gaussian mixture models Λ_α:

$$\begin{aligned} \hat{\alpha} &= h(X, \Lambda_\alpha) \\ &= \arg\max_{\alpha} p(X|\Lambda_\alpha) \end{aligned} \tag{6.16}$$

2. Adapted recognition pass as in two-pass VTN (cf. Eqn. 6.10)

Note that in this case no transformation of acoustic vectors is involved during training of the Gaussian mixture models, and when the models are applied for warping factor estimation in test.

The first approach of Wegmann et al. and Welling et al. has the advantage that the Gaussian mixture model is trained on the whole training corpus. In test, arbitrary

warping factors can be chosen to maximize the likelihood.

In the second approach of Lee and Rose, each model is trained on a different subset of the training corpus. It may happen that certain speakers with especially poor or good acoustic conditions will cause their corresponding Gaussian mixture models to give systematically lower or higher likelihoods. In early development tests it was found that warping factor estimates are more robust when the variance of all models is pooled. That is, each Gaussian mixture model Λ_α consists of a number of densities with different mean vectors $\mu_{\alpha,l}$ and mixture weight $c_{\alpha,l}$, but all vectors share the same diagonal covariance matrix Σ:

$$\begin{aligned} p(x|\Lambda_\alpha) &= \sum_{l=1}^{L} c_{\alpha,l} \cdot \mathcal{N}(x|\mu_{\alpha,l}, \Sigma; \Lambda_\alpha) \\ &\cong \max_l \{c_{\alpha,l} \cdot \mathcal{N}(x|\mu_{\alpha,l}, \Sigma; \Lambda_\alpha)\} \end{aligned} \tag{6.17}$$

Another minor disadvantage of the second approach is that a Gaussian mixture model can only be estimated reliably for those warping factors which occur frequently enough in the training data. Often the warping factor range has to be limited as there occur too few speakers with very small or large warping factors in training. Welling found, however, that the restriction of the warping factor range had no measurable impact on the word error rate in his tests [Welling 99]. This result has been confirmed on the corpora considered in this work, where two-pass recognition with a limited warping factor range $0.88 \leq \alpha \leq 1.12$ yielded the same word error rate as recognition with an extended range $0.80 \leq \alpha \leq 1.20$.

A major disadvantage of the first approach is that the warping factor is estimated by likelihood comparison of acoustic vector sequences normalized with different warping factors (cf. Eqn. 6.14). The Jacobian determinant (cf. Eqn. 2.12) is omitted during training of the Gaussian mixture model $\tilde{\Lambda}$ and during its application in test, which may cause systematic estimation errors [Pitz & Molau$^+$ 01]. The problem does not occur in the second approach, because in this case the acoustic vector sequence is not transformed. In the adapted recognition pass (Eqn. 6.10) the Jacobian can be neglected, as all acoustic vectors are normalized with the same warping factor $\hat{\alpha}$. Hence, the likelihoods of competing word sequences are all affected in the same way.

Another minor disadvantage of the first approach is a larger computational overhead as depicted in Figure 6.4. The signal analysis steps after the Fourier transform have to be repeated several times, since acoustic vectors normalized with all considered warping factors are required. In the second approach, these steps need to be carried out only twice (for the unwarped spectrum and for the best warping factor).

Recognition test results on the VerbMobil II corpus for both fast VTN approaches with different Gaussian mixture model complexities are summarized in Table 6.2. The first approach of Wegmann et al. and Welling et al. gives some improvements in recognition accuracy, but falls clearly short of the text-dependent two-pass VTN. Welling reported to train 64 densities for the Gaussian mixture model [Welling 99]. For the VerbMobil II

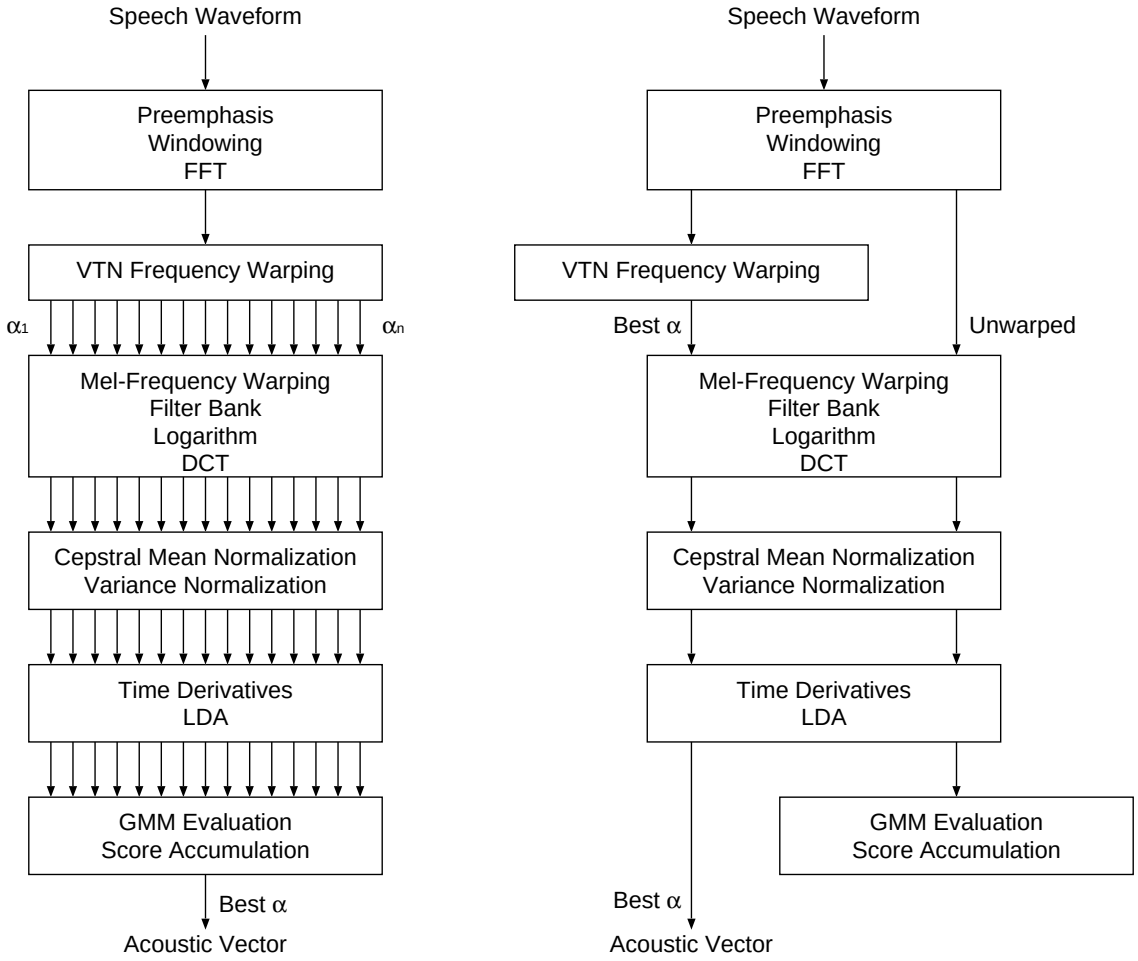

Figure 6.4: Comparison of the signal analyses for different text-independent warping factor estimation techniques (left: one Gaussian mixture model $\tilde{\Lambda}$ for normalized speech, right: Gaussian mixture models Λ_α for each warping factor α)

corpus, this number is somewhat small. If the model consists of 128 or more densities, the warping factors are estimated more reliable and the word error rate further decreases. This is consistent with Wegmann et al. who reported to use a 256 density Gaussian mixture model [Wegmann & McAllaster$^+$ 96].

The second approach of Lee and Rose improved by a pooled variance vector outperforms the first approach. It yields a word error rate similar to two-pass VTN. In addition, the warping factors are reliably estimated even if fewer densities per Gaussian mixture model are trained. The same number of densities per model as in the first approach means that the overall number of model parameters is higher (e.g. for 13 considered warping factors 13*128 Gaussian densities instead of 128). However, there is no increase in the number of likelihood calculations, as for each warping factor and each time frame the acoustic vector distance to all densities of one Gaussian mixture model has to be computed (e.g. 128 distance calculations per warping factor and time frame).

Table 6.2: Recognition test results on the VerbMobil II DEV99B corpus for different text-independent warping factor estimation schemes in test. Given are the the number of densities per Gaussian mixture model, the average number of active states after histogram pruning, the real-time factor, and the word error rate.

Fast VTN Warping Factor Estimation	# Densities per GMM	Search Space States	Search Space RTF	Overall [%] Del - Ins	Overall [%] WER
baseline without normalization		5792	13.4	4.1 - 4.7	24.9
one global GMM $\tilde{\Lambda}$ for normalized data	32	5617	13.2	4.5 - 5.0	24.6
	64	5472	13.0	4.4 - 5.0	24.5
	128	5374	12.9	4.3 - 5.1	24.1
	256	5337	12.9	4.3 - 4.9	24.0
GMMs Λ_α for each warping factor α	32	5180	12.7	3.7 - 4.9	22.9
	64	5194	12.7	3.8 - 4.7	**22.6**
	128	5130	12.7	3.8 - 4.9	22.7
	256	5170	12.7	3.9 - 4.8	22.9

With respect to the real-time factor, both approaches perform better than the baseline even though there was some computational overhead due to warping factor estimation. The reason is a reduction of the search space by up to 10% at identical pruning settings. It underlines that normalized acoustic models are more discriminant so that pruning becomes more efficient. As expected, the approach of Lee and Rose is somewhat faster than the one of Wegmann et al. and Welling et al., since signal analysis has to be carried out twice only (Figure 6.4). Whereas the real-time factor is constant for the latter approach, it *decreases* in the first with growing densities numbers of the Gaussian mixture model $\tilde{\Lambda}$. This surprising result has two reasons:

- The recognizer used quantized references and parallelized fast likelihood calculations (cf. Section 1.5). Hence, the number of distance calculations had only a limited impact on the overall real-time factor of the off-line system with conservative pruning settings.
- Warping factor estimation improved with increasing number of densities leading to a smaller number of active states during search. This more than leveled out the increased number of distance calculations for warping factor estimation.

Given the figures in Table 6.2 the conclusion could be drawn than the complexity of the Gaussian mixture model has a negligible impact on the real-time factor. For systems working at real-time as presented in the next section, however, likelihood calculations make up for a significant amount of computation time. In this case, tighter pruning settings to compensate for increased computational overhead during warping factor estimation result immediately in measurable performance degradation.

In all subsequent experiments, the second approach of Lee and Rose with 128 densities per Gaussian mixture model and a pooled covariance matrix was used for fast warping

factor estimation. The density number was chosen because it gave superior performance on a variety of corpora.

6.3.4 Incremental Warping Factor Estimation

In online recognition tasks it is desirable to have only little delay between speech recording and the output of the recognition result. This does not only require a fast recognizer, but also a signal analysis with minimum delay between recording and the start of search. So far, the warping factor was estimated sentence-wise on the whole test utterance before the recognition started. Text-independent warping factor estimation as presented in the previous section, however, does allow for an incremental warping factor estimation without additional delay (Figure 6.5).

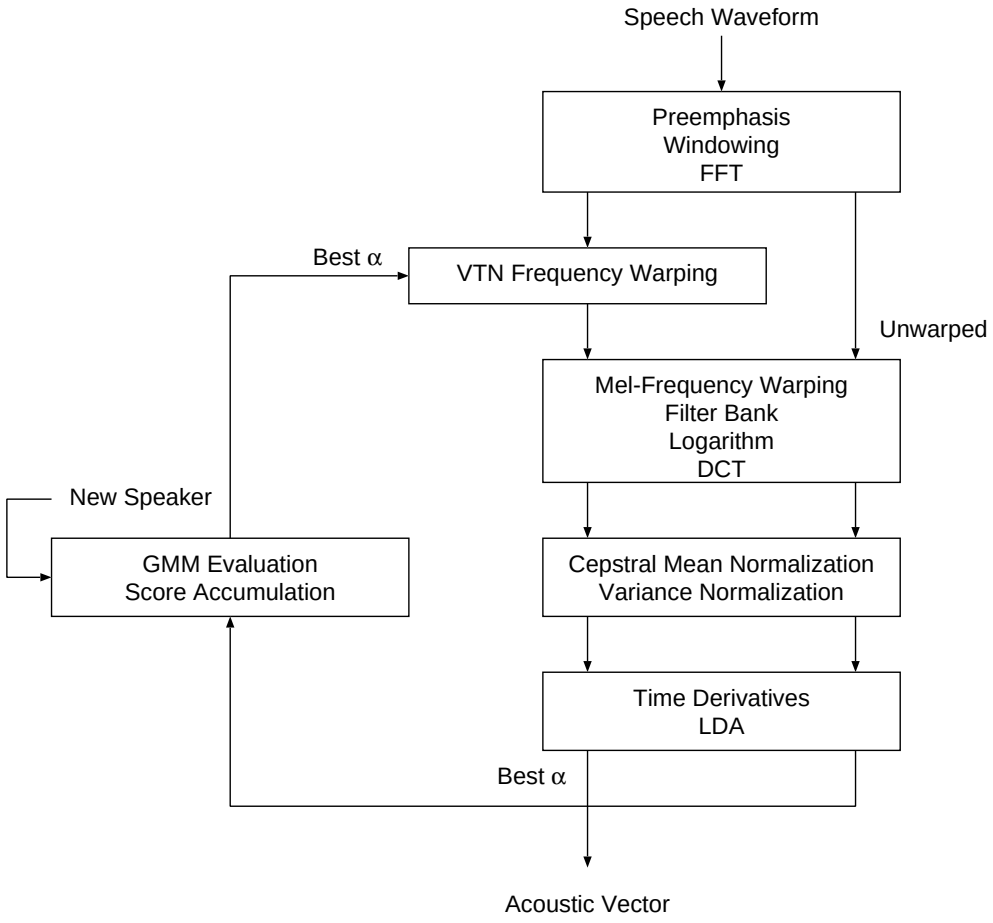

Figure 6.5: Signal analysis for fast VTN with incremental warping factor estimation.

The signal analysis steps after the Fourier transform are carried out twice. The acoustic vectors normalized with the currently best warping factor are used for recognition. The unnormalized acoustic vectors are immediately evaluated with the Gaussian mixture models and the probabilities $p(X|\Lambda_\alpha)$ are accumulated over time. Once a speaker change is detected (e.g. at the end of an utterance), the probabilities are reset and warping factor estimation starts from scratch again.

Problematic may be the first few acoustic vectors of a new speaker, when the warping factor has not yet settled. In this case, unnormalized acoustic vectors could be used in recognition for a predefined initialization time.

Table 6.3 summarizes recognition test results for fast vocal tract length normalization with sentence-wise warping factor estimation, and for incremental warping factor estimation with different initialization times. During initialization, the warping factor was forced to be 1.0, i.e. the vectors remained unnormalized. In general, sentence-wise warping factor estimation is more robust as it is based on more time frames. An initialization for incremental estimation is not necessary, as the word error rate changed only marginally when the estimated warping factor was used right from the beginning. In fact, there was a tendency to larger word error rates for increasing initialization times, since an ever growing part of each sentence remained unnormalized.

In Table 6.4, recognition test results are reported for two systems which were accelerated to almost real-time by a number of acceleration techniques (cf. Section 1.5) described in [Sixtus & Molau$^+$ 00] and [Kanthak & Sixtus$^+$ 00]. Fast vocal tract length normalization with incremental warping factor estimation was successfully applied in the RWTH speech recognition system used in the final VerbMobil II evaluation [Kanthak & Sixtus$^+$ 00]. It proves that vocal tract length normalization can significantly reduce the word error rate in online applications.

Table 6.3: Recognition test results on the VerbMobil II DEV99B corpus for sentence-wise and incremental warping factor estimation with different initialization times.

Fast VTN Warping Factor Estimation	Overall [%] Del - Ins	WER
sentence-wise	3.8 - 4.9	22.7
incremental without initialization	3.8 - 5.2	**23.3**
incremental with 1s initialization	4.0 - 4.9	23.4
incremental with 2s initialization	4.0 - 5.0	23.5
incremental with 3s initialization	4.0 - 5.0	23.5

Table 6.4: Recognition test results on the full VerbMobil II DEV99AB corpus for two systems accelerated to almost real-time. Given are the average number of active states after histogram pruning, the real-time factor, and the word error rate. The baseline system without normalization is compared to a VTN system with incremental warping factor estimation.

System	Search Space		Overall [%]	
	States	RTF	Del - Ins	WER
baseline without normalization	1839	1.3	5.8 - 3.9	25.1
fast VTN / incremental warping factor estimation	1771	1.2	5.5 - 3.5	**23.5**

6.4 Optimizations

Whereas efficient warping factor estimation schemes have been established at this stage, there are still a number of aspects that need to be considered to obtain best performance by VTN. Some of these will be discussed in the next section. In the end, the complete procedure for optimized vocal tract length normalization in training and test is summarized.

6.4.1 Frame Weighting

Wegmann et al. and Welling et al. reported that only speech frames should be used for the estimation of warping factors [Wegmann & McAllaster$^+$ 96] [Welling & Haeb-Umbach$^+$ 98]. Silence frames contain no information about the vocal tract length of the speaker, so they may only disturb the estimation. In the case of text-dependent warping factor estimation it is easy to omit silence frames, as there is an alignment between acoustic vectors and HMM states (cf. Figure 1.3). For the text-independent approach, Wegmann et al. used a "harmonicity feature" whereas Welling et al. applied the heuristics that the Gaussian mixture density with most observations is the "silence" density [Welling & Kanthak$^+$ 99]. All acoustic vectors which are closest to that density are regarded as silence vectors and subsequently omitted in the probability accumulation.

A negative side effect of leaving some times frames out is that the number of acoustic vectors used for warping factor estimation varies for different warping factors. The more the evaluated deviates from the true warping factor, the larger is the acoustic mismatch. Consequently, more speech frames are aligned to the silence mixture or density and do not contribute to the average sentence score used for warping factor estimation, i.e. to the negative log-likelihood of the sentence divided by the number of time frames. In practice, this discontinuity has the effect that especially in the case of short sentences the score as a function of the warping factor may not have a well defined minimum, which leads to some scatter in the estimated warping factors. An example is shown in Figure 6.6. The score over all 762 time frames of a test sentence from the VerbMobil II corpus has a clear minimum at $\alpha = 1.04$, but the warping factor estimated on speech frames only is ill-defined ($\alpha = 1.12$) because of their variable number.

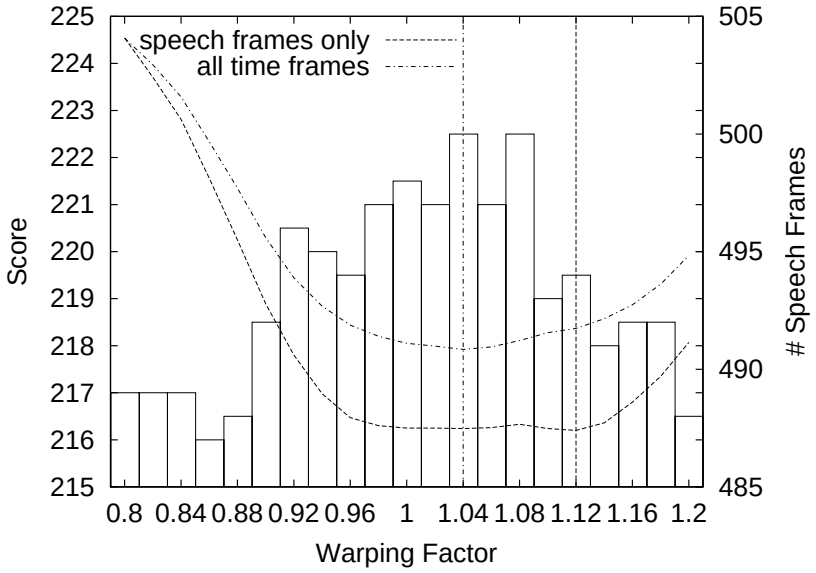

Figure 6.6: Average negative log-likelihood (score) of one sentence from the VerbMobil II corpus as a function of the warping factor (dashed lines). The number of speech frames as a function of the warping factor is plotted with bars. The best warping factors are marked with a vertical line.

To overcome this limitation, two alternative methods of frame weighting were evaluated. Instead of omitting silence frames, each acoustic vector is weighted by a factor $z(x)$ related to its energy (Eqn. 6.18). This technique boosts speech frames in a similar way as the method described above, but avoids discontinuities and the related problems. It also avoids mismatch since this method can be used in training and test, whereas before different criteria for the definition of a silence frame were used for text-dependent warping factor estimation in training and text-independent estimation in test.

As weight $z(x)$, the frames' energy $e(x)$ estimated on the magnitude spectrum (Eqn. 6.19) and the zeroth cepstrum coefficient $c_0(x)$, which is proportional to the logarithm of the frame energy (Eqn. 6.20), were used:

$$p(X|\Lambda) = \sum_{t=1}^{T} z(x_t) \cdot p(x_t|\Lambda) \tag{6.18}$$

$$e(x) = \frac{1}{N/2} \sum_{n=0}^{N/2-1} |x(e^{2\pi j \frac{n}{N}})| \tag{6.19}$$

$$c_0(x) = \frac{1}{N} \sum_{n=0}^{N/2-1} \lg |x(e^{2\pi j \frac{n}{N}})| \tag{6.20}$$

Recognition test results are summarized in Table 6.5. Both weights work equally well here, but additional tests on other corpora have shown that using the frames' energy $e(x)$ gives marginally better results. In fact, on this corpus it slightly hurt to omit silence frames by the heuristic method of Welling, but the differences were all very small.

Table 6.5: Recognition test results on the VerbMobil II DEV99B corpus for warping factor estimation on all frames, on speech frames only, and for frame weighting with the frames' energy $e(x)$ and the zeroth cepstrum coefficient $c_0(x)$.

Fast VTN Warping Factor Estimation	Overall [%] Del - Ins	WER
using all frames	3.8 - 4.9	22.5
omitting silence frames	3.8 - 4.9	22.7
frame weight with $e(x)$	3.9 - 4.6	**22.3**
frame weight with $c_0(x)$	3.8 - 4.8	22.3

6.4.2 Warping Functions

The model of the vocal tract introduced in Section 6.1 predicts a linear shift of formant frequencies depending on the length of the vocal tract. A direct implementation of linear frequency axis warping is difficult, however, due to the limited bandwidth of the acoustic signal.

Telephone audio data are typically sampled at 8 kHz and microphone data at 16 kHz. The Nyquist theorem states that frequencies larger than half the sampling frequency cannot be reconstructed from the sampled signal, i.e. the bandwidth is limited to 4 kHz for telephone and 8 kHz for microphone data. Linear warping of a line spectrum changes the bandwidth [Lee & Rose 96]: It either requires higher frequencies beyond the bandwidth (for $\alpha < 1$) or the highest frequencies are discarded (for $\alpha > 1$), which results in information loss. Hence, a piece-wise linear warping function is often applied. Up to a limiting frequency ω_0 of 3.5 / 7 kHz (telephone and microphone data, respectively) at the unwarped frequency axis, the spectra are warped linearly with the warping factor α. Above that limiting frequency they are warped with another factor which is chosen such that the bandwidth remains unchanged (Eqn. 6.21 and Figure 6.3) [Wegmann & McAllaster$^+$ 96][Welling & Haeb-Umbach$^+$ 98]:

$$\begin{aligned} \omega \rightarrow \tilde{\omega} &= f_\alpha(\omega) \\ &= \begin{cases} \alpha \cdot \omega & \omega \leq \omega_0 \\ \omega_0 + \frac{\pi - \omega_0}{\pi - \omega_0/\alpha} \cdot (\omega - \omega_0/\alpha) & \omega > \omega_0 \end{cases} \qquad (6.21) \\ \omega_0 &= \frac{7}{8}\pi \qquad (6.22) \end{aligned}$$

The warping function has an upper limit of $\alpha = 8/7 \cong 1.14$, since for larger warping factors the bandwidth is exceeded before the limiting frequency ω_0 is reached. For some male speakers, however, warping factors as large as 1.2 are observed (cf. Figure 6.11). For this reason, the turning frequency was re-defined in a symmetric fashion to be 3.5 / 7 kHz at the unwarped frequency axis for $\alpha \leq 1$, and at the warped frequency axis for $\alpha > 1$ (Eqn. 6.23 and Figure 6.7, left). Uebel and Woodland also applied a symmetric piece-wise linear function in their tests [Uebel & Woodland 99]:

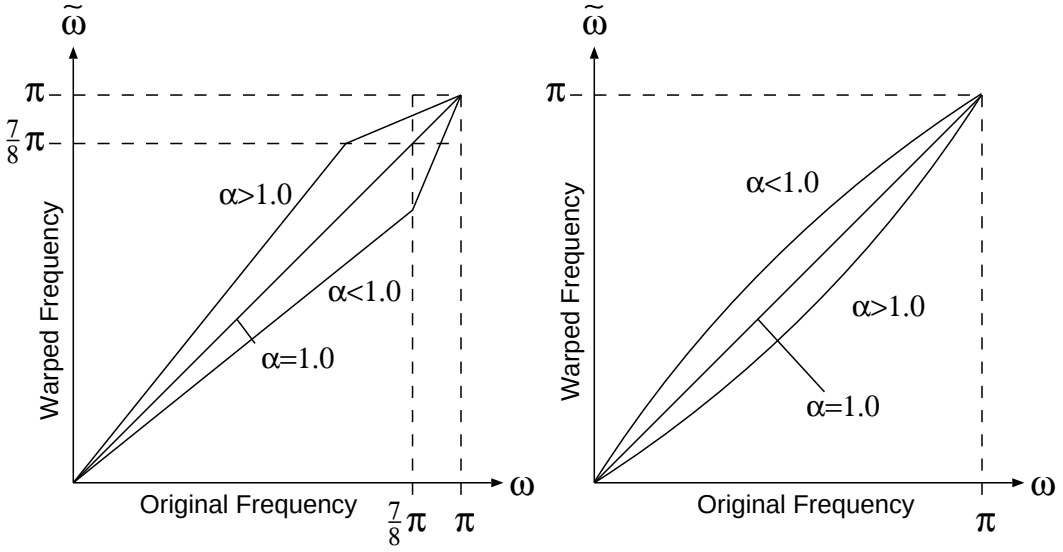

Figure 6.7: Schematic plot of a symmetric piece-wise linear (left) and power (right) frequency axis warping function.

$$\omega_0 = \begin{cases} \frac{7}{8}\pi & \alpha \leq 1 \\ \\ \frac{7}{8 \cdot \alpha}\pi & \alpha > 1 \end{cases} \tag{6.23}$$

The modification had no measurable impact on the word error rate, but the warping factor range could now be extended beyond $\alpha = 1.14$ to avoid possible side effects from a limited warping factor range.

As an alternative to piece-wise linear warping with a turning point, a power warping function was tested (Eqn. 6.24 and Figure 6.7, right). It is not perfectly consistent with the linear shift of formant frequencies predicted by the simple tube model (cf. Section 6.1), but it is reasonably close to linear frequency axis warping and has no discontinuity:

$$\begin{aligned} \omega \rightarrow \tilde{\omega} &= f_\alpha(\omega) \\ &= \left(\frac{\omega}{\pi}\right)^\alpha \cdot \pi \end{aligned} \tag{6.24}$$

Recognition test results for both functions are presented in Table 6.6.

Table 6.6: Recognition test results on the VerbMobil II DEV99B corpus for different frequency axis warping functions.

Two-Pass VTN	Overall [%]	
Warping Function	Del - Ins	WER
symmetric piece-wise linear	5.0 - 3.9	**23.9**
power function	5.2 - 4.1	24.3

There are only minor differences in the word error rate, but the piece-wise linear function performs slightly better [Molau & Kanthak$^+$ 00], which confirms the findings of other groups (cf. Section 3.1.1).

6.4.3 Re-estimation of CART and LDA

Vocal tract length normalization removes undesired variations from the speech signal. Formants are shifted to reference positions which results in more discriminant acoustic vectors as shown by the superior recognition performance. It is to expect that re-estimation of the phonetic decision tree (CART) and the LDA transformation matrix on normalized training data improves the acoustic modeling and further reduces the word error rate.

In initial re-estimation tests, no gain in recognition accuracy was observed, which was in accordance with earlier results of Welling [Welling 99]. After system optimization, however, a minor but consistent performance improvement was obtained on all investigated corpora. Table 6.7 gives the result for the full VerbMobil II DEV99AB test corpus. The same improvement of 0.4% absolute was obtained on the same corpus with an across-word VTN system [Sixtus 02]. The missing improvements in earlier systems might have been caused by sub-optimal procedures for warping factor estimation and acoustic model training.

6.4.4 Iterative Warping Factor Estimation

In Section 6.3.3 a warping factor estimation scheme for test speakers was introduced that is based on a set of Gaussian mixture models Λ_α. It avoids systematic errors possibly introduced by neglecting the Jacobian determinant, because there is no transformation of the acoustic vector involved. The problem still persists at an earlier stage, however, as the individual Gaussian mixture models are trained according to warping factors of the training speakers (cf. Eqn. 6.15). These were derived by a procedure where the Jacobian should have been taken into account: The training data are normalized with different warping factors, and the one that maximizes the likelihood is chosen (cf. Eqn. 6.2).

Table 6.7: Recognition test results on the full VerbMobil II DEV99AB corpus for re-estimation of the phonetic decision tree and the LDA transformation matrix on normalized training data.

Two-Pass VTN	Overall [%]	
Estimation of CART and LDA	Del - Ins	WER
baseline without normalization	5.1 - 4.4	25.7
on unnormalized training data	4.9 - 4.0	23.8
on normalized training data	5.4 - 3.7	**23.4**

It was explained in Section 6.2 that the acoustic model θ_0 used to determine the warping factors of training speakers should at be a normalized model. Since at the beginning there are no normalized data available, an unnormalized single density model is used instead. Once the warping factors are estimated, however, a new single density model $\tilde{\theta}_0$ may be trained on normalized data, and warping factor estimation may be repeated in an iterative fashion:

$$\tilde{\theta}_0 \quad \cong \quad \arg\max_{\theta} \prod_{r=1}^{R} \tilde{p}(X_r^{\hat{\alpha}_r} | W_r; \theta) \tag{6.25}$$

This approach is similar to the iterative training procedure of Lee and Rose [Lee & Rose 96]. However, the training corpus is not split and the acoustic model is of low resolution, which significantly reduces the computation time.

Recognition test results for re-estimated warping factors on the VerbMobil II corpus are summarized in Table 6.8. A marginal improvement is observed for one or two extra iterations of warping factor estimation. Unfortunately, the improvements were not consistently obtained on all corpora. One additional iteration was helpful in all cases when the original decision tree and LDA transformation matrix were used. However, when the decision tree and the LDA transformation matrix were re-estimated on normalized data and the warping factor estimation was repeated, the word error rate deteriorated slightly on some corpora. This is consistent with earlier results of Lee and Rose [Lee & Rose 96], who also reported increasing word error rates on test data for more than one iteration of warping factor estimation. Thus, warping factors are typically not re-estimated because of substantially higher computational costs and a negligible gain in recognition performance at best.

Table 6.8: Recognition test results on the full VerbMobil II DEV99AB corpus for iterative estimation of warping factors in training.

Two-Pass VTN		Overall [%]	
Single Density Model θ_0 Trained on	CART/LDA Trained on	Del - Ins	WER
unnormalized data (first iteration)	unnormalized data	4.9 - 4.0	23.8
normalized data (second iteration)		5.3 - 3.7	23.6
normalized data (third iteration)		5.2 - 3.9	23.9
unnormalized data (first iteration)	normalized data	5.4 - 3.7	23.4
normalized data (second iteration)		5.3 - 3.5	23.5
normalized data (third iteration)		5.4 - 3.5	**23.3**

6.5 Conclusions

Given the experimental results presented in the previous sections, the following baseline procedure for vocal tract length normalization starting from an existing system without VTN is proposed to achieve maximum recognition performance:

1. Train a low-resolution (single density) acoustic model θ_0 on unnormalized training data X (cf. Eqn. 6.3).

2. Estimate warping factors $\hat{\alpha}_r$ for the training speakers r using θ_0 and the reference transcription (cf. Eqn. 6.2).

3. Normalize the training data with the calculated warping factors (cf. Eqn. 6.4).

4. Re-estimate the decision tree and LDA transformation matrix on the normalized training data $\tilde{X}$.

5. For two-pass VTN, train an unnormalized acoustic model θ on unnormalized data X with the new decision tree and LDA transformation matrix for the first recognition pass (cf. Eqn. 6.8).

6. For fast warping factor estimation (cf. Eqn. 6.16), train Gaussian mixture models Λ_α (cf. Eqn. 6.15) for each warping factor α on all unnormalized training data X_r from speakers r with warping factor $\alpha_r = \alpha$. Use the new LDA transformation matrix.

7. Train a normalized acoustic model $\tilde{\theta}$ (cf. Eqn. 6.5) on the normalized data $\tilde{X}$ with the new decision tree and LDA transformation matrix for the final adaptive recognition pass (cf. Eqn. 6.10).

A symmetric piece-wise linear function is applied for spectral warping (cf. Eqn. 6.21 and 6.23) with warping factors ranging from 0.80 to 1.20 in steps of 0.02. All acoustic vectors are considered for warping factor estimation in training and test, but weighted with their energy (cf. Eqn. 6.18 and 6.19).

6.6 Final Results for Different Corpora

Using the procedure proposed in the previous section, vocal tract length normalization was applied to different large-vocabulary speech corpora, namely VerbMobil II, North American Business News 20k, and EuTrans II. Results for within-word systems are summarized in Table 6.9, and for across-word systems in Table 6.10. For each task, the word error rate of optimized baseline systems without VTN is given as reference.

Fast VTN achieves in general about the same word error rate as two-pass VTN, which means that the full gain in recognition performance can be realized without an increase of the real-time factor. The reduction in word error rate ranges between 8% and 9% relative for the within-word systems, which is of the same order as the improvement

Table 6.9: Within-word system recognition test results for different large vocabulary corpora and different acoustic conditions. Given are word error rates for optimized baseline systems without vocal tract length normalization, and for two-pass and fast VTN.

Corpus	VTN	Overall [%]	
		Del - Ins	WER
VerbMobil II DEV99AB	baseline without VTN	5.1 - 4.4	25.7
	fast	5.4 - 3.7	23.5
	two-pass	5.4 - 3.5	**23.3**
NAB 20k	baseline without VTN	1.5 - 2.2	12.5
	fast	1.5 - 2.2	**11.5**
	two-pass	1.4 - 2.2	11.6
EuTrans II	baseline without VTN	4.2 - 3.1	16.5
	fast	3.5 - 3.4	15.3
	two-pass	3.9 - 2.9	**15.1**

obtained by across-word models alone.

A combination of vocal tract length normalization and across-word models achieves relative word error rate reductions between 11% and 16% compared to the baseline within-word system without normalization, i.e. the gain in recognition performance by both techniques is not fully additive. In a complex speech recognition system, a differentiation between errors caused by inappropriate vocal tract or coarticulation modeling is not possible. The only interpretation is that some of the recognition errors that are avoided by normalization are also not made by across-word modeling, and vice versa.

Table 6.10: Across-word system recognition test results for different large vocabulary corpora and different acoustic conditions. Given are word error rates for optimized baseline systems without vocal tract length normalization, and for two-pass and fast VTN.

Corpus	VTN	Overall [%]	
		Del - Ins	WER
VerbMobil II DEV99AB	baseline without VTN	5.3 - 3.6	23.3
	fast	4.3 - 3.4	**21.4**
	two-pass	4.6 - 3.3	21.6
NAB 20k	baseline without VTN	1.4 - 2.0	11.5
	fast	1.3 - 2.2	11.0
	two-pass	1.5 - 2.2	**10.9**
EuTrans II	baseline without VTN	3.9 - 3.2	15.7
	fast	4.2 - 2.4	**14.7**
	two-pass	4.2 - 2.7	15.0

On the EuTrans II corpus, across-word models yielded only a comparatively small improvement in recognition performance. A possible explanation is the small training corpus of only 8 hours of speech data. There are probably too few across-word contexts observed during training to get a similar gain as on the large training corpora. In this case, warping factors were estimated in a speaker-incremental fashion, i.e. all previous sentences of a speaker were used for warping factor estimation in addition to the current sentence. This had essentially no effect on the within-word results, but slightly improved the across-word results for VTN.

6.7 Integrated Frequency Axis Warping

In this section, a novel concept to derive Mel-frequency cepstral coefficients (MFCCs) directly from the magnitude spectrum of the speech signal will be introduced and analyzed. A number of successive steps of the traditional signal analysis including VTN frequency axis warping are integrated into the cepstrum transformation, which avoids possible quantization and interpolation errors. The same idea of merging successive signal analysis step after the Fourier transform into a single step was proposed by Yu and Waibel [Yu & Waibel 00], but they followed a completely different approach.

6.7.1 Motivation

The signal analysis front-end of a speech recognition system was described in Section 1.2. Here we concentrate on the steps between the Fourier and the cosine transform. Every 10 ms, the Fourier transform is applied to a short segment of the speech signal which yields a spectrum with 512 spectral lines in the RWTH system. The magnitude spectrum is warped according to the Mel-scale [Davis & Mermelstein 80] in order to adapt the frequency resolution to the properties of the human ear. Independently of this psycho-acoustic explanation it was shown that the Mel-scale is the optimal choice to reduce the spectral resolution at high frequencies [Mashao 96]. Then, the spectrum is segmented into a number of critical bands by means of a filter bank, which typically consists of overlapping triangular filters. The dynamic range of the filter bank coefficients is reduced by taking the logarithm, and the discrete cosine transform is applied to get raw MFCC vectors.

Mel-frequency warping and the filter bank can be implemented easily in the frequency domain (Figure 6.8). One method is to transform the magnitude spectrum, i.e. to compute a Mel-warped spectrum by interpolation from the original discrete-frequency magnitude spectrum [Wegmann & McAllaster$^+$ 96]. The advantage is that the following triangular filters all have the same shape and can be placed uniformly at the Mel-warped spectrum. However, the quantization may be critical due to the large dynamic range of the magnitude spectrum.

Another way is to place the triangular filters non-uniformly at the unwarped spectrum [Lee & Rose 96] and thereby implicitly incorporate Mel-frequency scaling. However, quantization errors may occur if the spectral resolution is not appropriate. The lowest

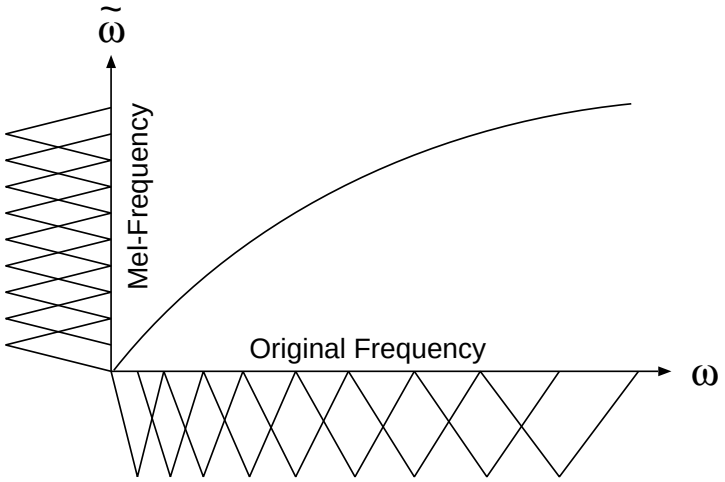

Figure 6.8: Schematic plot of different triangular filter bank implementations for Mel-frequency warping. The filters are either uniformly distributed at the Mel-warped spectrum, or non-uniformly at the original spectrum.

filters could be placed at a very few spectral lines only, and the maximum of one of the filters may fall just in-between two spectral lines. In addition, the filters should not be triangular and symmetric anymore, but bend according to the shape of the Mel-function at the position of the filter. Finally, this approach is problematic in the case of vocal tract length normalization with linear frequency axis warping, as the bandwidth changes which may place the highest filters beyond the Nyquist frequency [Lee & Rose 96][Chu & Jie$^+$ 97].

Last but not least, it is not clear how many filters are required and which filter shape is optimal. Triangular filters are occasionally replaced by trapezoidal or more complex shaped ones derived from auditory models.

An alternative first presented in [Molau & Pitz$^+$ 01a] is to omit the filter bank and compute cepstral coefficients directly on the log-magnitude spectrum. It avoids possible problems of the standard approaches by integrating spectral warping into the discrete cosine transform. A comparison of the traditional signal analysis with the integrated approach proposed here is shown in Figure 6.9.

6.7.2 Integration of Frequency Axis Warping into DCT

Ignoring spectral warping for a moment, cepstral coefficients c_k are defined as follows:

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega \lg |X(e^{j\omega})| \cdot e^{j\omega k} \tag{6.26}$$

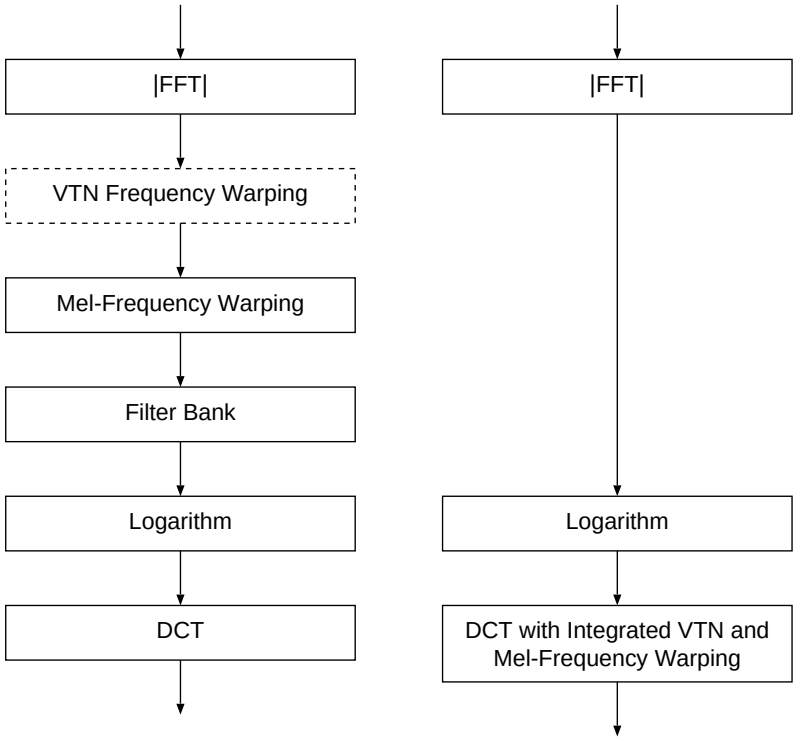

Figure 6.9: Comparison of the traditional MFCC signal analysis with the integrated frequency axis warping approach.

Depending on whether or not a filter bank is used, $|X(\cdot)|$ denotes either the filter bank coefficients or the magnitude spectrum.

The sequential application of a monotone invertible frequency axis warping function $g:\ [-\pi,\pi] \to [-\pi,\pi]$ and the discrete cosine transform can be expressed as follows:

$$\begin{aligned} \omega &\rightarrow \tilde{\omega} = g(\omega) \\ c_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} d\tilde{\omega} \lg |X(e^{jg^{-1}(\tilde{\omega})})| \cdot e^{j\tilde{\omega}k} \end{aligned} \tag{6.27}$$

To incorporate frequency axis warping into the cosine transform, we change the integration variable and apply the derivative of the warping function $d\tilde{\omega}/d\omega$ (Eqn. 6.28). The continuous integral is then approximated in the standard way by a discrete sum (Eqn. 6.29):

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega \lg |X(e^{j\omega})| \cdot e^{jg(\omega)k} \cdot g'(\omega) \tag{6.28}$$

$$\cong \frac{1}{N} \sum_{n=0}^{N/2-1} \left\{ \lg |X(e^{2\pi j \frac{n}{N}})| \cdot \cos[g(2\pi n/N) \cdot k] \cdot g'(2\pi n/N) \right\} \tag{6.29}$$

Eqn. 6.29 describes how frequency axis warping can be integrated in general into the cosine transform. It leads to a compact implementation of the vector of warped cepstrum coefficients c with only a few lines of code. The transformation reduces to a matrix multiplication (Eqn. 6.30) of a transformation matrix U (Eqn. 6.31) and the log-magnitude spectrum m. Since U depends only on the fixed frequency axis warping function $g(\omega)$ and its derivative, it can be computed beforehand:

$$c \cong U \cdot m \qquad c, m \in \mathbb{R}^{N/2} \qquad U \in \mathbb{R}^{N/2 \times N/2} \tag{6.30}$$

$$\begin{aligned} c &= \left(c_0, c_1, \ldots, c_{N/2-1}\right) \\ m &= \left(\lg|X(e^{2\pi j \frac{0}{N}})|, \lg|X(e^{2\pi j \frac{1}{N}})|, \ldots, \lg|X(e^{2\pi j \frac{N/2-1}{N}})|\right) \\ U_{i,j} &= \cos\{g(2\pi i/N) \cdot j\} \cdot g'(2\pi i/N) \qquad i,j = 0, \ldots, N/2-1 \end{aligned} \tag{6.31}$$

Specific equations for Mel-frequency warping and VTN warping as well as recognition test results will be given in the following two sections.

6.7.3 Integration of Mel-Frequency Warping

Mel-frequency warping $\mu(\omega)$ is usually carried out according to Eqn. 6.32 (adapted from [Young 93]) with f_s denoting the sampling frequency:

$$\begin{aligned} \omega \rightarrow \tilde{\omega} &= \mu(\omega) \\ &= 2595 \cdot \lg\left(1 + \frac{\omega f_s}{2\pi \cdot 700Hz}\right) \end{aligned} \tag{6.32}$$

For integration into the cosine transform, the Mel-warping function needs to be normalized (Eqn. 6.33) to meet the criterion $\tilde{\mu}(\pi) = \pi$. In addition, the derivative $\tilde{\mu}'(\omega)$ is required (Eqn. 6.34):

$$\begin{aligned} \tilde{\mu}(\omega) &= \pi \cdot \frac{\mu(\omega)}{\mu(\pi)} \\ &= d \cdot \lg\left(1 + \frac{\omega f_s}{2\pi \cdot 700Hz}\right) \\ d &= \frac{\pi}{\lg\left(1 + \frac{f_s}{2 \cdot 700Hz}\right)} \end{aligned} \tag{6.33}$$

$$\tilde{\mu}'(\omega) = \frac{d \cdot f_s}{(2\pi \cdot 700Hz + \omega \cdot f_s) \cdot \ln(10)} \tag{6.34}$$

Replacing $g(\omega)$ and $g'(\omega)$ in Eqn. 6.29 by $\tilde{\mu}(\omega)$ and $\tilde{\mu}'(\omega)$ yields the desired transformation matrix for Mel-frequency cepstral coefficients.

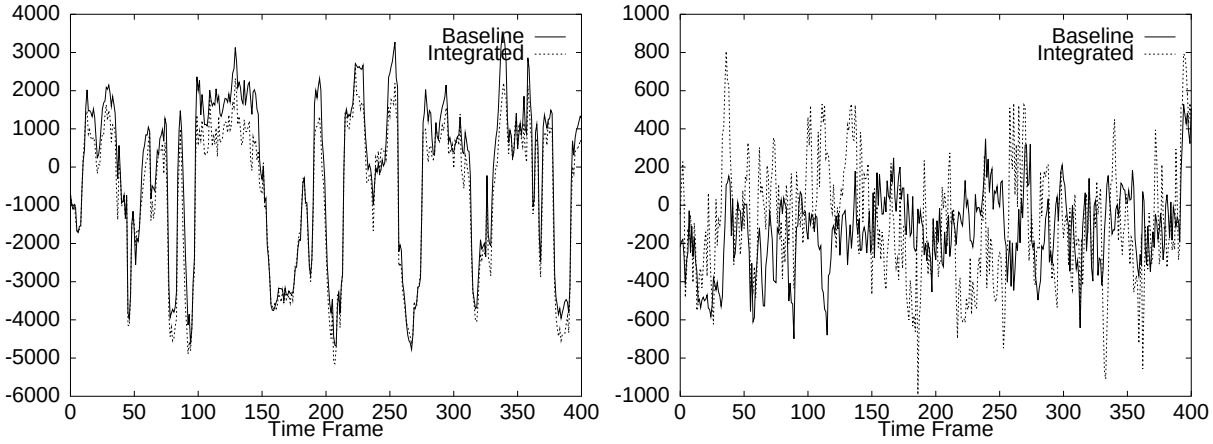

Figure 6.10: Comparison of cepstrum coefficients 1 (left) and 15 (right) for a test sentence from the VerbMobil II corpus. Depicted are the traditional filter bank approach and the cosine transform with integrated Mel-frequency warping.

Figure 6.10 shows the effect of the modified signal analysis on two cepstrum coefficients for a test sentence from the VerbMobil II corpus. Whereas the lower order coefficients are almost identical, the differences increase for higher coefficient. Main reason is the missing filter bank, which reduces the spectral resolution before the cosine transform.

Recognition test results for integrated Mel-scaling on the VerbMobil II corpus are summarized in Table 6.11. The integrated approach is not only coequal to the baseline approach with filter bank, but even yields marginally better results on this corpus.

6.7.4 Integration of VTN Frequency Warping

Vocal tract length normalization is like Mel-frequency warping a technique that relies on frequency axis warping of the magnitude spectrum. One possible implementation of VTN is to modify the location of filters in the filter bank (cf. Figure 6.8) just as for Mel-frequency scaling [Lee & Rose 96]. Another method that is applied in the RWTH speech recognition system is to compute a warped spectrum by interpolation from the original discrete-frequency magnitude spectrum [Wegmann & McAllaster$^+$ 96][Welling 99].

Table 6.11: Recognition test results on the full VerbMobil II DEV99AB corpus for different Mel-frequency warping methods. Results are given for the baseline filter bank approach, and for the cosine transform with integrated Mel-frequency warping.

Mel-frequency Warping	Overall [%] Del - Ins	WER
baseline	4.9 - 4.8	25.7
integrated	5.0 - 4.4	**25.3**

From the equations presented in Section 6.7.2 it is clear that VTN frequency axis warping can also be fully integrated into the cepstrum transformation.

The frequency axis warping function $\nu_\alpha : [0, \pi] \rightarrow [0, \pi]$ for vocal tract length normalization needs to be monotone and invertible. As discussed in Section 6.4.2, this holds for all typical VTN warping functions in order to prevent loss of information.

To avoid complicated case distinctions for different warping factors and frequencies, we re-write the symmetric piece-wise linear warping function (cf. Eqn. 6.21) in the following convenient form:

$$\begin{aligned} \omega \rightarrow \tilde{\omega} &= \nu_\alpha(\omega) \\ &= \beta_\omega \omega + \kappa_\omega \end{aligned} \tag{6.35}$$

The parameters β_ω and κ_ω depend formally on ω, but in practice they can take on two values only (with the limiting frequency ω_0 as defined in Eqn. 6.23):

$$\beta_\omega = \begin{cases} \alpha & \omega \leq \omega_0 \\ \frac{\pi - \alpha \cdot \omega_0}{\pi - \omega_0} & \omega > \omega_0 \end{cases} \tag{6.36}$$

$$\kappa_\omega = \begin{cases} 0 & \omega \leq \omega_0 \\ (\alpha - 1) \cdot \frac{\pi \cdot \omega_0}{\pi - \omega_0} & \omega > \omega_0 \end{cases} \tag{6.37}$$

Mel-warping is applied after the magnitude spectrum is warped by vocal tract length normalization. Hence, the combination $\chi(\omega)$ of VTN and Mel-frequency warping (Eqn. 6.38) and its derivative (Eqn. 6.39) become:

$$\begin{aligned} \chi(\omega) &= \tilde{\mu}(\nu_\alpha(\omega)) \\ &= d \cdot \lg\left(1 + \frac{\{\beta_\omega \omega + \kappa_\omega\} \cdot f_s}{2\pi \cdot 700Hz}\right) \end{aligned} \tag{6.38}$$

$$\chi'(\omega) = \frac{d \cdot \beta_\omega \cdot f_s}{(2\pi \cdot 700Hz + \{\beta_\omega \omega + \kappa_\omega\} \cdot f_s) \cdot \ln(10)} \tag{6.39}$$

Cepstrum coefficients with integrated VTN and Mel-frequency warping are obtained by replacing $g(\omega)$ and $g'(\omega)$ in Eqn. 6.29 by $\chi(\omega)$ and $\chi'(\omega)$. In this case, and own transformation matrix U (cf. Eqn. 6.31) has to be calculated for each warping factor. As the number of considered warping factor is typically limited (e.g. to 21 discrete values in the range $0.80 \leq \alpha \leq 1.20$), the transformation matrices can still be computed beforehand. During normalization, the correct matrix for frequency axis warping is selected.

The integration of spectral warping into the cosine transform lead to an interesting observation. When the filter bank was omitted and the integrated approach was used to estimate the warping factors of the training speakers, their distribution became smoother than before. Figure 6.11 shows the corresponding histograms of warping factors for

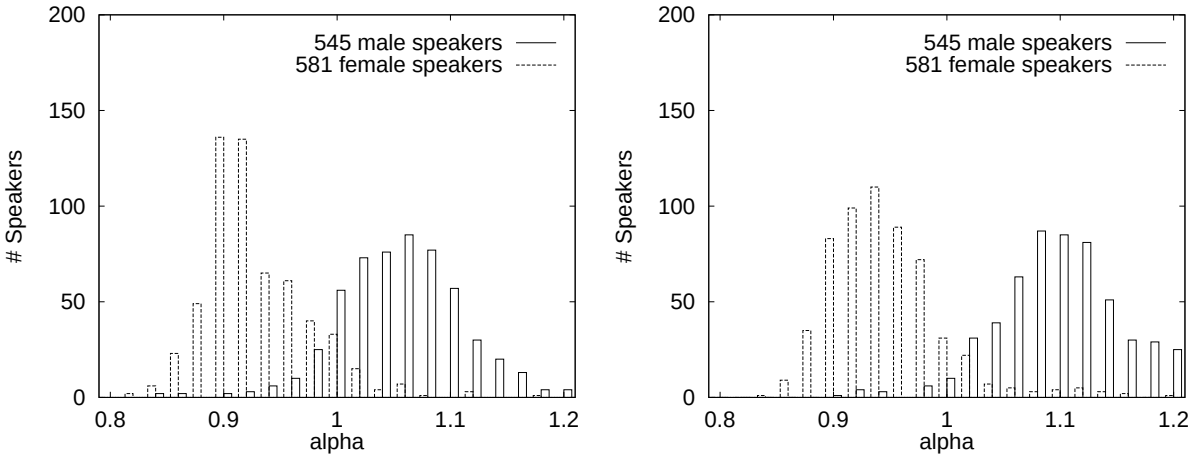

Figure 6.11: Warping factor distribution of the VerbMobil II training speakers. The left histogram was obtained in the traditional way with filter bank and frequency axis warping by linear interpolation of the discrete-frequency spectra, the right histogram with integrated VTN and Mel-frequency warping.

male and female speaker in the VerbMobil II training corpus. The difference is even more prominent if the histograms are compared with the old distributions reported by Welling [Welling & Kanthak$^+$ 99]. A closer inspection revealed that linear interpolation of spectral lines during VTN frequency axis warping was one of the reasons for the uneven distribution observed before. A similar smooth warping factor histogram could be obtained by the traditional approach when the spectral lines were interpolated in the logarithmic domain. Another reason was the initial beam size used for the forced alignment in warping factor estimation (cf. Eqn. 6.2), which was too small. It turned out, however, that the word error rate was unaffected by the improved distribution.

Recognition test results for the full VerbMobil II DEV99AB corpus are summarized in Table 6.12. Both with fast and two-pass vocal tract length normalization, the integrated approach performs similar to the baseline approach. The results reported here were obtained without decision tree and LDA transformation matrix re-estimation on normalized data, which is why the baseline recognition results differs from those reported earlier.

Table 6.12: Recognition test results on the full VerbMobil II DEV99AB corpus for the traditional and the integrated VTN and Mel-frequency warping approach.

Spectral Warping	VTN	Overall [%] Del - Ins	WER
baseline	fast	4.5 - 4.5	23.8
integrated		5.0 - 4.1	24.0
baseline	two-pass	4.4 - 4.3	**23.8**
integrated		4.9 - 4.1	24.0

6.7.5 Improved Spectral Smoothing

One advantage of integrated frequency axis warping is a better control over the amount of spectral smoothing. The standard triangular filter bank reduces the spectral resolution from the original 512 spectral lines in the RWTH signal analysis to typically 15/20 filter bank channels (telephone/microphone data). The cosine transform yields at most the same number of Mel-frequency cepstral coefficients. Using all coefficients makes no sense, however, because it would reduce the cepstrum transformation to a plain linear transformation. After linear discriminant analysis, the resulting acoustic vector would be identical with or without the cosine transform. In practice, only the first 12/16 cepstral coefficients (telephone/microphone data) are typically calculated and used, which has an additional spectral smoothing effect.

When the filter bank is omitted and spectral warping as well as the cosine transform are applied to the log-magnitude spectrum, the spectral resolution is not yet reduced. The number of cepstrum coefficients computed and used for further processing is the only factor that controls the amount of spectral smoothing. It is possible to derive a larger number of cepstral coefficients and preserve more spectral information, for example.

Figure 6.12 illustrates these effects. The left side shows the traditional approach and the right side the integrated frequency axis warping approach without filter bank. The magnitude spectrum of a vowel was taken from an utterance of the VerbMobil II corpus and warped according to the Mel-frequency (solid line). In the standard approach, it was first processed by a 20 channel filter bank. The filter bank coefficients are symbolized by the horizontal lines (left). They preserve some of the early spectral peaks, but most of the pitch signature is removed. Next, the cosine transform was calculated to derive 16 Mel-frequency cepstral coefficients. The inverse cosine transform was applied to these coefficients for demonstration purposes yielding the smoothed spectrum (dashed line).

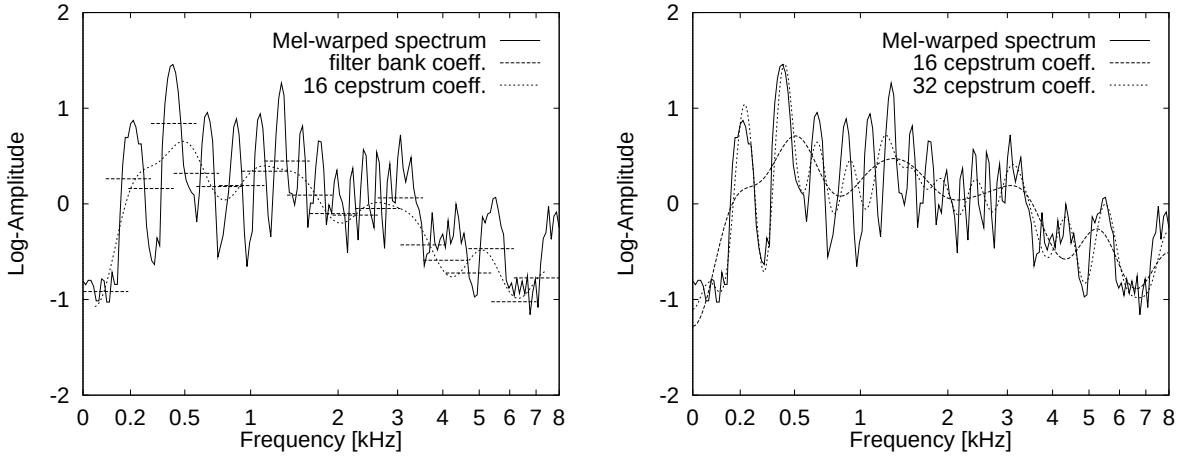

Figure 6.12: Comparison of spectral smoothing by the traditional signal analysis with filter bank (left) and by the integrated frequency axis warping approach (right). The amplitude is shown in a logarithmic scale. Details are given in the text.

The fine structure of the original Mel-warped spectrum is completely removed, what remains is the energy distribution in different spectral bands.

In the integrated approach on the right side, 16 Mel-frequency cepstral coefficients were derived directly from the logarithm of the unwarped magnitude spectrum. The inverse cosine transform was applied again to visualize the smoothed spectrum, which looks similar to the smoothed filter bank spectrum. By increasing the number of cepstral coefficients, more spectral information is preserved. Doubling the number of cepstrum coefficients maintains the first peaks of the pitch, whereas fine structures at higher frequencies are still smoothed out.

In the RWTH signal analysis, the cepstrum vector is further processed by mean (and possibly variance) normalization (cf. Figure 1.2). It is augmented by time derivatives, and a number of successive cepstrum vectors are transformed by linear discriminant analysis, which reduces the size of the acoustic vector. Hence, by increasing the number of cepstral coefficients it is left to the LDA transformation to find those coefficients and linear combinations of them that discriminate best between the different LDA classes. This makes integrated frequency axis warping more similar to the approach of Yu and Waibel, who apply only one unified linear transformation to derive the reduced acoustic vector from the log-magnitude spectrum [Yu & Waibel 00].

Table 6.13 summarizes recognition test results on the full VerbMobil II DEV99AB corpus with a variable degree of spectral smoothing. The number of cepstrum coefficients taken from the integrated approach was increased starting at 16, which is identical to the number of cepstral coefficients in the traditional signal analysis. The raw cepstrum vector was augmented with the full first derivatives, and the second derivative of the energy. Three consecutive augmented cepstrum vectors were fed into the LDA transformation. In order to avoid side effects, the dimensionality of the acoustic vector after the LDA transformation was kept fixed at the original value of 33. The optimal number of cepstrum coefficients was found to be 64, which gave a clear reduction in word error rate over the baseline system.

It can be concluded that cepstral coefficients above 16 still contain information that is relevant for the speech recognition process. Even if much of the original spectral fine structure including the pitch is preserved when as many as 64 cepstral coefficients are calculated, linear discriminant analysis is able to extract the important pieces of information. It is to expect that at a certain point LDA parameter estimation becomes a serious problem, though. In the above setup with 64 cepstral coefficients, the full LDA transformation matrix was of dimension $\{(64+64+1)*3\}^2$, i.e. 387×387. The number of matrix elements to be estimated from data increased by a factor of 15 compared to the baseline case with 16 cepstral coefficients.

Note that the baseline result could not be improved by simply doubling the number of filter bank channels and cepstral coefficients (second line in Table 6.13). The reason are probably increasing quantization errors especially of the lower filters when the bandwidth of each filter is halved.

Table 6.13: Recognition test results on the full VerbMobil II DEV99AB corpus for integrated spectral warping with an increasing numbers of cepstral coefficients, and with an enlarged acoustic vector. Results for the traditional approach are gives as the baseline.

Spectral Warping	# MFCC	Dimensionality of the Acoustic Vector	Overall [%] Del - Ins	WER
baseline	16	33	4.9 - 4.8	25.7
	32		5.2 - 4.5	25.9
integrated	16		5.0 - 4.4	25.3
	32		5.0 - 4.6	25.4
	48		5.2 - 4.5	25.1
	64		5.2 - 4.2	**24.9**
	80		5.1 - 4.4	25.2
baseline	16	48	5.0 - 4.4	25.2
integrated	64		4.6 - 4.8	**24.3**

In another experiment the size of the LDA-transformed vector was increased by 50%, as more spectral information might require a larger acoustic vector for optimum performance. The word error rate indeed further decreased by half a percent (Table 6.13). However, the same performance gain was obtained when the size of the acoustic vector of the baseline system was increased by the same amount. In both cases, the computation time for training and test increased significantly.

6.7.6 Results for Different Corpora

Integrated spectral warping was tested on two different large-vocabulary speech recognition corpora with and without vocal tract length normalization. The results for VerbMobil II were presented in Tables 6.11, 6.12 and 6.13. If the number of cepstrum coefficients is left unchanged, the recognition performance is similar to the traditional signal analysis approach. Omitting the filter bank and integrating Mel-frequency warping into the cepstrum transformation simplifies the signal analysis (no filter bank parameters need to be optimized), avoids possible interpolation and quantization problems, and leads to a more compact implementation of the MFCC front-end. Concepts like vocal tract length normalization that rely on warping the frequency axis can be easily integrated as well. When the number of cepstrum coefficients was increased, a 4% relative reduction of the word error rate could be achieved.

Results for the North American Business News corpus are summarized in Table 6.14. Note that the VTN results reported here were obtained without decision tree and LDA transformation matrix re-estimation, which is why the baseline results differ from those reported earlier. The presented approach yields again the same word error rate as the traditional signal analysis. However, increasing the number of cepstrum coefficients did not further improve the recognition accuracy in this case. As the NAB corpus consists of

Table 6.14: Recognition test results on the NAB 20k corpus for the traditional and the integrated VTN and Mel-frequency warping approach.

Spectral Warping	# MFCC	VTN	Overall [%] Del - Ins	WER
baseline	16	-	1.5 - 2.3	12.5
integrated	16		1.5 - 2.3	**12.4**
integrated	64		1.6 - 2.2	12.5
baseline	16	fast	1.4 - 2.3	11.9
integrated			1.5 - 2.2	11.8
baseline	16	two-pass	1.4 - 2.4	11.8
integrated			1.4 - 2.2	**11.7**

read speech, it lacks spontaneous speech phenomena that make the recognition task more difficult. It might be concluded that fewer spectral information is sufficient to characterize clean read speech.

6.8 Summary

Vocal tract length normalization is a model based technique that aims at reducing inter-speaker variations of mean formant frequencies by warping the frequency axis during signal analysis.

The baseline procedure of warping factor estimation in training and test was introduced, and a number of optimizations and improvements were implemented and tested. Piecewise linear frequency axis warping turned out to be superior to non-linear warping. Weighting of acoustic vectors during warping factor estimation, and re-estimation of the phonetic decision tree and the LDA transformation matrix proved to be helpful, whereas iterative warping factor estimation in training yielded only a negligible gain at best.

The optimized vocal tract length normalization scheme in training and test yielded consistently large improvements of 8% to 9% relative by two-pass recognition on all corpora under investigation. In addition it was shown that the full gain in recognition accuracy can be obtained without an increase in computation time. This was achieved by text-independent warping factor estimation based on a Gaussian mixture models trained on unnormalized data. The requirements of online recognition were met by incremental warping factor estimation.

A novel integrated frequency axis warping approach was developed that merges a number of successive signal analysis steps into a single one. The filter bank can be omitted, the logarithm is applied directly to the spectral lines, and all frequency axis warping schemes are integrated into the cepstrum transformation. The approach avoids possible quantization and interpolation problems of other techniques and yields a compact implementation of

Mel-frequency cepstral coefficients by a simple matrix multiplication of the log-magnitude spectrum. It was shown that integrated frequency warping yielded the same recognition performance as the traditional approach with filter bank. In addition, it allows for a better control over the amount of spectral smoothing. Increasing the number of cepstral coefficients without enlarging the acoustic vector improved the recognition performance on the VerbMobil II corpus.

Chapter 7

Histogram Normalization and Rotation

7.1 Histogram Normalization

In this chapter, a data distribution based normalization technique (cf. Section 2.4) will be studied. The idea of histogram normalization will be introduced, the basic normalization scheme will be evaluated, and a number of extensions will be proposed that increase the overall gain in recognition performance.

7.1.1 Principle

Histogram normalization (or histogram equalization) is a widely used technique in image processing, object recognition and computer vision (e.g. [Ballard & Brown 82], pp. 70–71), but there have been only few applications in speech recognition so far.

The principal idea is as follows: Suppose the training and test data are distributed as depicted in the two-dimensional example feature space in Figure 7.1. There is a mismatch between both data sets that may have different reasons as discussed in Section 2.1. The mismatch will be especially prominent if there are major differences in the recording environments. Histogram normalization transforms the test to the training data distribution by mapping the depicted marginal distributions [Dharanipragada & Padmanabhan 00]. In the generalized approach presented here, both training and test data are mapped to some pre-defined reference distribution.

Histogram normalization relies on two basic assumptions:

1. The global statistics of the speech signal are independent of what is actually spoken, i.e. the phoneme frequencies in training and test are similar.
2. The feature space dimensions are oriented such that the variations that are tackled by histogram normalization are uncorrelated in each dimension.

Under these conditions, each feature space dimension can be mapped independently of the others - a significant simplification.

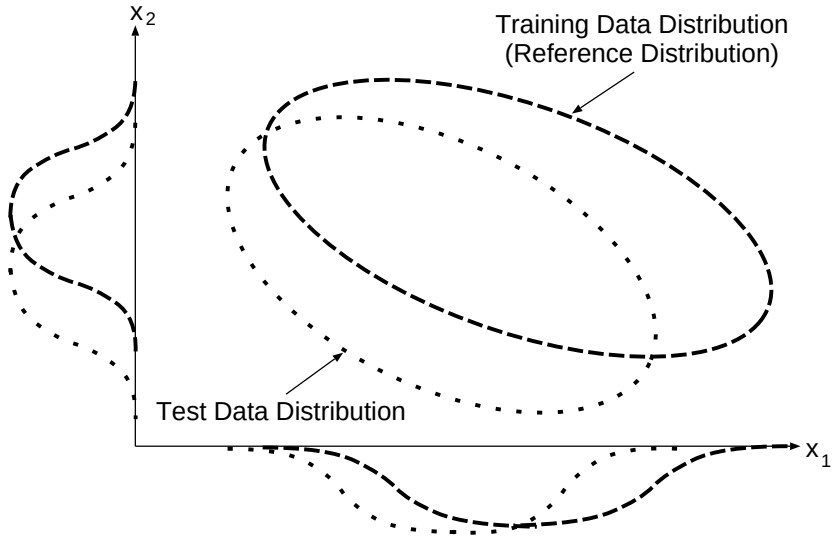

Figure 7.1: Schematic distribution of training and test data in a two-dimensional example feature space. The marginal distributions are plotted along both axes.

The basic normalization algorithm is as follows: First, the reference histogram to which all data is mapped has to be defined. Usually the overall distribution of the training data is used for reference. This choice is somewhat arbitrary, as various other distributions could be used as well. It can be argued, however, that the inherent distribution of the training data is a good choice to start with.
For each feature space dimension (note that for notational simplicity the dimension index is omitted in all equations):

1. Compute a normalized histogram $\tilde{p}(x)$ on the full training corpus.
2. Compute the cumulative training data histogram $\tilde{P}(x)$ which becomes the reference histogram:

$$\tilde{P}(x) = \int_{-\infty}^{x} dx' \tilde{p}(x') \tag{7.1}$$

In the normalization step, the parameter set α_r (Eqn. 2.9) of the transformation function $f_\alpha(x)$ (Eqn. 2.11) has to be determined for each condition (cf. Section 2.3). In the case of histogram normalization, the condition-dependent distributions $p_r(x)$ and $P_r(x)$ are calculated. For each condition $r = 1, \ldots, R$ and each feature space dimension:

3. Compute a normalized histogram $p_r(x)$ from all data X_r.
4. Compute the cumulative condition-dependent histogram $P_r(x)$:

$$P_r(x) = \int_{-\infty}^{x} dx' p_r(x') \tag{7.2}$$

Finally, the transformation is applied to all data X_r from condition r:

5. Replace each value x by $\tilde{x}$ that corresponds to the same point in the cumulative reference histogram (Figure 7.2):

$$\begin{aligned} x \to \tilde{x} &= f_\alpha(x) \\ P_r(x) &\stackrel{!}{=} \tilde{P}(\tilde{x}) \\ \tilde{x} &= \tilde{P}^{-1}(P_r(x)) \end{aligned} \tag{7.3}$$

Since the normalization depends on the acoustic data only, it amounts to an additional signal analysis step that is independent of training and test. From the transformed training data a normalized acoustic model $\tilde{\theta}$ is derived (cf. Eqn. 2.10), and the transformed test data are used for recognition.

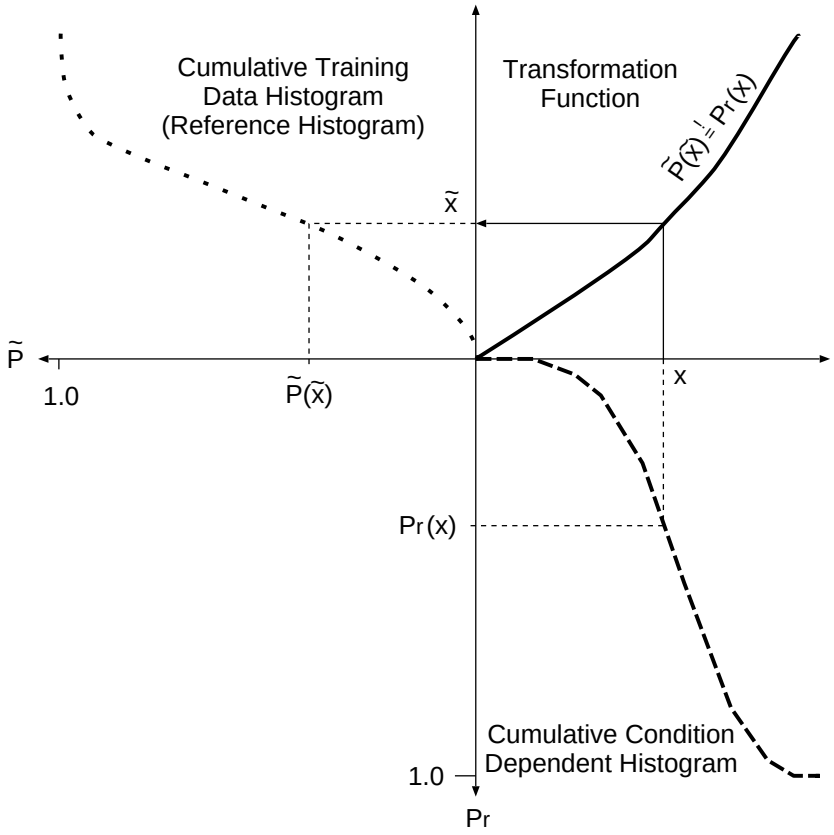

Figure 7.2: Principle of histogram normalization: Data X_r from condition r are transformed such that the cumulative condition-dependent histogram $P_r(x)$ matches the cumulative reference histogram $\tilde{P}(\tilde{x})$.

Histogram normalization has a number of convenient properties:

- it is text-independent and relies only on global statistics of the speech data
- it is a non-parametric, discrete approximation of a complex non-linear transformation function and makes no assumption about the functional form of the transformation
- once the histograms are calculated, histogram normalization can be implemented by a simple table-lookup, so it is also computationally attractive

Histogram normalization can account for scaling, shifting, or any type of non-linear distortion of each feature space dimension but, due to the assumption of uncorrelated features, not for possible feature space rotations. In the case depicted in Figure 7.1, basic histogram normalization will reduce the mismatch significantly but not remove it completely, because the feature space is rotated by a small amount.

7.1.2 Definition of the Acoustic Conditions

An important aspect of histogram normalization is the definition of the acoustic conditions r, for which a particular histogram $P_r(x)$ is estimated. The definition is task-dependent and has to meet the following requirements:

- there has to be enough data for each condition (typically one or more minutes) to estimate the histogram reliably
- each condition should contain data for only a single speaker in order to allow for the normalization of possible speaker-dependent variations in the speech signal
- the channel conditions should be constant to allow for the normalization of possible channel-dependent distortions

Estimating one histogram on the full test corpus meets the first requirement but violates the other two, whereas sentence-wise normalization would meet only the latter two requirements but not the first. Hence, in the following analyses, histogram normalization is applied either turn-wise, i.e. a condition contains all utterances from one speaker in one conversation (VerbMobil II), or speaker-wise, if all data from a speaker were collected under identical channel conditions (EuTrans II, CarNavigation). The average amount of data available for estimating the condition-dependent histograms is listed under "average condition duration" in the corpus descriptions (Chapter 5).

The first requirement prevents the use of histogram normalization in on-line recognition tasks or on small data samples. There are two solutions if only a few seconds worth of adaptation data are available: Either a coarse histogram with fewer bins and appropriate interpolation in-between is estimated, or a parametric transformation function is applied whose parameters are estimated from histogram statistics. These approaches have been investigated in detail by Padmanabhan and Dharanipragada [Padmanabhan & Dharanipragada 01] and by Hilger et al. [Hilger & Ney 01][Hilger & Molau$^+$ 02] and are not pursued further in this work.

7.1.3 Histogram Normalization in Test only

In previous work, histogram normalization has been applied in test only. This is a special case of the generalized approach presented here. The overall distribution of the training data is used for reference as well, but only the test data are mapped to the reference histogram. The data from the individual training conditions and therefore also the acoustic model remain unnormalized.

In Section 2.2 a theoretic explanation was given why normalization of the test data alone results often in moderate gain of recognition performance only, whereas full performance is achieved when both test and training data are normalized. Corresponding recognition test results are summarized in Table 7.1. Results are presented for the normalization at different signal analysis stages (discussed in more detail in the following section). As expected, the best results on the VerbMobil II corpus are obtained when both training and test data are normalized.

7.1.4 Normalization Stages

Dharanipragada and Padmanabhan proposed a normalization of cepstral features [Dharanipragada & Padmanabhan 00]. There are, however, a number of stages in the signal analysis front-end where histogram normalization may be applied (cf. Figure 2.4):

- In the course of signal analysis, the speech waveform is transformed into a sequence of spectra by means of a Fourier transform. Each individual spectral line could be regarded as an independent distribution that needs to be normalized. For computational reasons it is more practical to apply histogram normalization after the filter bank, which leaves typically 15 or 20 (telephone or microphone data) distributions for the normalization. As the logarithm is a monotone function, it makes no difference whether histogram normalization is applied before or after the logarithm. In practice, spectral log compression before normalization helps to keep quantization errors small.

Table 7.1: Recognition test results on the VerbMobil II DEV99B corpus for basic histogram normalization with and without training data normalization.

Histogram Normalization		Overall [%]	
Stage	Training Data Norm.	Del - Ins	WER
baseline without normalization		4.9 - 4.4	24.6
log filter bank	no	5.0 - 4.0	23.8
	yes	4.9 - 4.4	**23.0**
cepstrum	no	4.5 - 4.3	24.0
	yes	5.0 - 4.7	24.3
after LDA	no	4.6 - 4.3	24.2
	yes	4.9 - 4.4	24.1

Histogram normalization of the log filter bank coefficients may help to reduce spectral distortions that are limited to certain frequency bands. It also normalizes the energy distribution in each frequency band.

- The mean of cepstral coefficients is typically subtracted in order to remove time-invariant channel transfer functions. In some tasks it also helps to scale cepstral coefficients to unity variance. Histogram normalization at the cepstrum stage, however, has a larger degree of freedom. It may not only shift and scale the distribution of each cepstral coefficient, but also distort it non-linearly.

- Linear discriminant analysis of cepstral coefficients and their time derivatives is a standard feature of the RWTH large vocabulary speech recognition system (cf. Section 1.2), since it consistently improves the recognition accuracy on all tasks [Welling 99]. The LDA-transformed vector is the one that is finally presented to the speech recognizer. Hence, applying histogram normalization after linear discriminant analysis will normalize the distribution of acoustic test vectors to that observed during training of the corresponding acoustic model.

In addition, it is possible to apply histogram normalization sequentially at different stages in a multi-pass scheme: After the reference distributions are defined, the condition-dependent histograms of the first normalization stage can be derived in a first signal analysis pass. In the next pass, the acoustic vectors can be normalized at the first stage, and the condition-dependent histograms for the second stage can be accumulated, etc. In the end, the distributions of the acoustic vector components can be normalized at all stages.

As it is a-priori unknown at which stage of signal analysis histogram normalization performs best, or if there is a gain by sequential normalization at different stages, all three stages and combinations were tested. The results for the VerbMobil II corpus are summarized in Table 7.2.

Table 7.2: Recognition test results on the VerbMobil II DEV99B corpus for basic histogram normalization at different signal analysis stages.

Histogram Normalization			Overall [%]	
Log Filter Bank	Cepstrum	after LDA	Del - Ins	WER
baseline without normalization			4.9 - 4.4	24.6
yes	no	no	4.9 - 4.4	23.0
no	yes	no	5.0 - 4.7	24.3
no	no	yes	4.9 - 4.4	24.1
yes	yes	no	4.9 - 4.4	22.9
yes	no	yes	4.3 - 4.0	**22.5**
no	yes	yes	4.9 - 4.2	24.0
yes	yes	yes	4.9 - 4.3	22.7

It turns out that histogram normalization performs well at the filter bank stage, and that there are only marginal improvements when normalization is performed on cepstrum or LDA-transformed vectors. A possible explanation is that most of the variations compensated for by histogram normalization are uncorrelated in the spectral domain. Note that not the individual filter bank channels are supposed to be uncorrelated, but that the spectral distortions seem to be restricted to certain frequency bands.

The performance improvement of histogram normalization at different stages is to some extent additive, but the computational effort increases significantly due to the multi-pass signal analysis.

7.1.5 Histogram Smoothing

As an example, Figure 7.3 (left) shows the reference histogram for the third log filter bank coefficient obtained on the VerbMobil II training corpus. It turns out that the distributions of most filter bank channels, cepstral coefficients, and LDA-transformed vector components have a similar bimodal shape.

The original distributions can be replaced by mixtures of two Gaussian densities as reference histogram (Figure 7.3, right) which smoothes data scatter efficiently and results in better modeling of outliers. The mixtures are fitted to the observed distributions with a least squared error criterion which better matches the tails of the distribution than maximum likelihood estimates.

Replacing the observed histograms by Gaussian mixtures as reference helps to improve the recognition accuracy as shown in Table 7.3 for the VerbMobil II corpus. Even though the lowest word error rate of 22.5% could not be reduced any further, this result is now obtained by normalizing the filter bank coefficients alone. The normalization is much faster and easier if cepstrum and LDA feature vector normalization can be omitted.

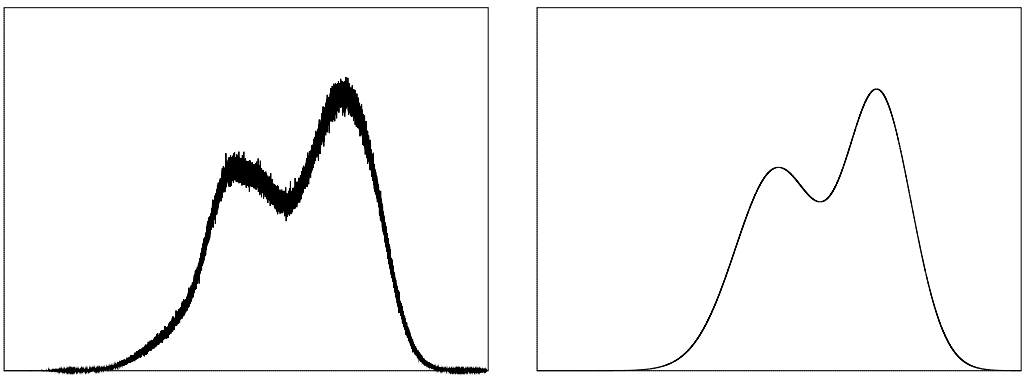

Figure 7.3: Reference histogram $\tilde{p}(x)$ for the third log filter bank coefficient obtained on the VerbMobil II training corpus. The bimodal distribution (left) can be well approximated by a Gaussian mixture with two densities (right).

Table 7.3: Recognition test results on the VerbMobil II DEV99B corpus for basic histogram normalization with a smoothed reference histogram.

Histogram Normalization			Overall [%]	
Log Filter Bank	Cepstrum	LDA	Del -Ins	WER
baseline without normalization			4.9 - 4.4	24.6
yes	no	no	4.6 - 3.8	**22.5**
no	yes	no	5.2 - 4.3	23.9
no	no	yes	4.9 - 4.2	23.7
yes	yes	no	4.9 - 4.1	23.2
yes	no	yes	4.7 - 3.9	22.8
no	yes	yes	4.8 - 4.4	23.8
yes	yes	yes	4.8 - 3.8	**22.5**

Hence, further histogram normalization tests have been carried out at the log filter bank stage only.

7.1.6 Silence Fraction Treatment

The first assumption of histogram normalization about the global statistics of the speech signal (cf. Section 7.1.1) is often violated. Even if enough speech data is available to ensure that the phoneme frequency is about the same for each condition, and even if the acoustic realization of the phonemes is identical, the histograms may still vary due to different silence fractions. This has a severe impact on conditions with a much lower or higher than average silence fraction. In the first case, histogram normalization will transform a number of acoustic speech vectors to silence and cause more deletions of words. In the latter case, some silence vectors will be transformed to speech and cause word insertions.

Figure 7.4 shows a histogram of the condition-wise silence fractions in the VerbMobil II corpus. Non-speech events like hesitations or transcribed noise items are considered as "speech" in this context. The average silence fraction is 17%, but the number varies between 3% and 76% for individual conditions.

Two possible solutions to the problem rely on having separate reference histograms for speech and silence. In the first solution, two streams of acoustic vectors are fed into the speech recognizer. One of them is adapted to the speech, the other to the silence histogram. During recognition it is known at each point in time, if the current state hypothesis belongs to speech or not. The corresponding acoustic vector can be chosen for likelihood calculations. A disadvantage of this approach is the discontinuity of the acoustic vectors at each speech/silence boundary introduced by the different reference histograms. The same problem occurs if a speech/silence detector is used prior to recognition.

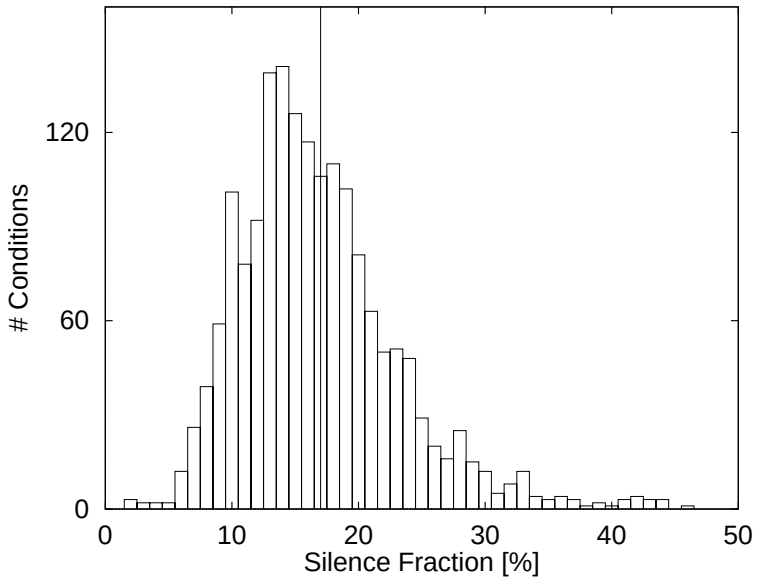

Figure 7.4: Histogram over the silence fractions γ_r of individual conditions r in the Verbmobil II training corpus. The vertical line marks the average silence fraction of 17%.

A conceptually simpler solution pursued here is to determine the silence fraction of each condition r beforehand and create condition-dependent reference histograms $\tilde{P}_r(x)$ from the speech and silence histogram that are adapted to the observed silence fraction. In this approach, the discontinuity is avoided and the speech recognizer needs no modifications.

To obtain the speech and silence histograms, a forced alignment with the reference transcription is carried out on the training data. All acoustic vectors mapped to the silence mixture are accumulated in the silence histogram $\tilde{p}_{sil}(x)$. All other vectors are accumulated in the speech histogram $\tilde{p}_{sp}(x)$. It can be seen that the bimodal structure of most histograms observed before (cf. Section 7.1.5) is in fact a manifestation of speech and silence. (Figure 7.5). The first peak can be almost completely attributed to silence frames, whereas the second peak is mainly caused by more energetic speech frames.

In the normalization step, the silence fraction γ_r of the actual training or test condition r has to be determined first. For the training data, it is estimated as before by forced alignment with the reference transcription. Since in test the correct transcription is unknown, the silence fraction has to be calculated either in a preliminary recognition pass (two-pass recognition) or with a dedicated speech/silence detector (e.g. as described in [Macherey & Ney 02]).

For each condition $r = 1, \ldots, R$, an adapted reference histogram $\tilde{P}_r(x)$ is computed by linear interpolation between the speech and silence histograms. Note that the same result is obtained whether the normalized histograms $\tilde{p}_{sil}$ and $\tilde{p}_{sp}$ are interpolated before the cumulative histogram is computed, or whether the cumulative histograms $\tilde{P}_{sil}$ and $\tilde{P}_{sp}$ are interpolated (Eqn. 7.4). The latter approach is computationally more efficient, though:

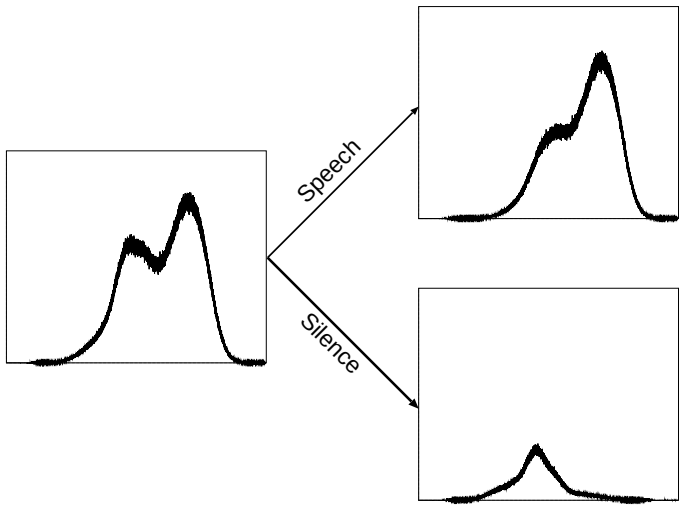

Figure 7.5: Histogram over the third log filter bank coefficient on the VerbMobil II training corpus. The left side shows the original reference histogram, on the right side the histogram is split into speech and silence. The speech and silence histograms are not yet normalized.

$$\tilde{P}_r(x) = \int_{-\infty}^{x} dx' \tilde{p}_r(x') = \gamma_r \cdot \tilde{P}_{sil}(x) + (1-\gamma_r) \cdot \tilde{P}_{sp}(x) \tag{7.4}$$

$$\tilde{p}_r(x) = \gamma_r \cdot \tilde{p}_{sil}(x) + (1-\gamma_r) \cdot \tilde{p}_{sp}(x) \tag{7.5}$$

$$\tilde{P}_{sil}(x) = \int_{-\infty}^{x} dx' \tilde{p}_{sil}(x') \qquad \tilde{P}_{sp}(x) = \int_{-\infty}^{x} dx' \tilde{p}_{sp}(x') \tag{7.6}$$

The adapted reference histograms $\tilde{P}_r(x)$ are used for normalization of training and test data as in the basic histogram normalization approach (cf. Section 7.1.1). As an example, Figure 7.6 shows the reference histogram of the third log filter bank coefficient for three different silence fractions. The left histogram adapted to a silence fraction

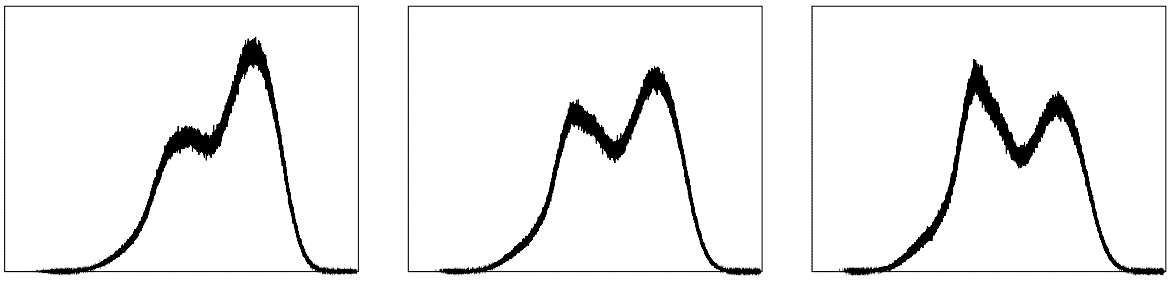

Figure 7.6: Reference histogram $\tilde{p}_r(x)$ for the third log filter bank coefficient on the VerbMobil II training corpus adapted to three different silence fractions γ_r.

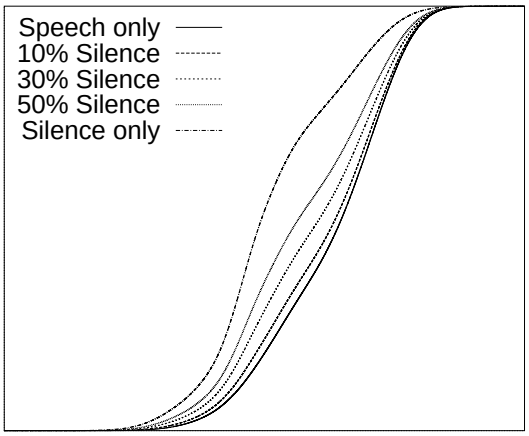

Figure 7.7: Cumulative reference histogram $\tilde{P}_r(x)$ for the third log filter bank coefficient on the VerbMobil II training corpus adapted to three different silence fractions γ_r.

of 10% is most similar to the original histogram for speech and silence (Figure 7.3 and Figure 7.5, left), because this value is closest to the average silence fraction of the training corpus. The larger the silence fraction, the more prominent becomes the first "silence" peak, whereas the second "speech" peak loses significance. The corresponding cumulative histograms for different silence fractions are depicted in Figure 7.7.

Recognition test results for histogram normalization of log filter bank coefficients are summarized in Table 7.4. They show that the treatment of the silence fraction helps to further improve the recognition performance. In fact, on the EuTrans II corpus, histogram normalization only improved the recognition accuracy in connection with silence fraction treatment (see Section 7.4).

In this and all further tests, histogram smoothing was not applied anymore. Informal test have shown that a smoothed reference histogram gave no further gain in recognition performance. A possible explanation is that the individual histograms of speech and silence deviate significantly from Gaussian distributions, which is why smoothing by Gaussian mixtures may introduce larger deviations from the original training data distribution.

Table 7.4: Recognition test results on the VerbMobil II DEV99B corpus for histogram normalization of log filter bank vectors with and without silence fraction treatment.

Histogram Normalization Silence Fraction Treatment	Overall [%] Del - Ins	 WER
baseline without normalization	4.9 - 4.4	24.6
no	4.6 - 3.8	23.0
yes	4.2 - 3.9	**21.8**

7.1.7 Histogram, Mean, and Variance Normalization

So far, histogram normalization has been regarded as a supplementary normalization step. The baseline signal analysis scheme was left unchanged and histogram normalization has been implemented as an additional module between log compression of the filter bank coefficients and the cepstrum transformation. Since mean and variance normalization are carried out at the cepstrum stage, they were implicitly applied after histogram normalization.

Since the cepstrum transformation is linear, mean normalization can also be carried out before the discrete cosine transform, i.e. at the same stage as histogram normalization. There is in fact a close link between these techniques: Mean normalization matches the means of the training and test data distributions, variance normalization transforms both distributions to have unity variance, and histogram normalization matches the overall shape of the training and test data distributions. Furthermore, the feature space dimension are treated independently of each other in all cases.

The question arises whether mean and variance normalization should be carried out before histogram normalization at the log filter bank stage, or after histogram normalization at the cepstrum stage. Corresponding recognition test results for the baseline system and for histogram normalization with silence fraction treatment are summarized in Table 7.5.

As expected, mean normalization alone yields identical word error rates independently at which stages it is applied (24.5% vs. 24.6%). This is also true for additional histogram normalization (23.6% vs. 23.5%).

Variance normalization at the log filter bank stage performs worse than at the cepstrum stage. Log filter bank variance normalization degrades the recognition performance significantly both with and without subsequent histogram normalization. The baseline

Table 7.5: Recognition test results on the VerbMobil II DEV99B corpus for mean and variance normalization of log filter bank and cepstrum coefficients in the baseline system, and in connection with histogram normalization.

Normalization Steps in the Order of their Application	Overall [%] Del - Ins	WER
log filter bank mean	4.6 - 5.0	24.5
log filter bank mean & variance	5.1 - 4.1	25.2
cepstral mean	4.4 - 4.9	24.6
cepstral mean & variance (baseline)	4.9 - 4.4	24.6
log filter bank mean, histogram	4.6 - 4.1	23.6
log filter bank mean & variance, histogram	5.0 - 5.2	26.0
histogram, cepstral mean	4.8 - 4.0	23.5
histogram, cepstral mean & variance	4.2 - 3.9	**21.8**

setup with cepstral variance and no histogram normalization yields the same word error rate of 24.6% as a system without variance normalization (contrary to earlier development tests when the baseline system was optimized). However, the performance clearly improves if cepstral variance normalization is applied after histogram normalization.

In summary, best results were achieved when mean and variance normalization were applied after histogram normalization at the cepstrum stage. This result was confirmed on other corpora. Dharanipragada and Padmanabhan performed cepstral mean after histogram normalization as well [Dharanipragada & Padmanabhan 00]. Note that histogram normalization transforms the data condition-wise, whereas mean and variance normalization are applied sentence-wise or in a sliding window depending on the speech corpus (cf. Chapter 5).

7.2 Feature Space Rotation

7.2.1 Motivation

The second basic assumption of histogram normalization is that the feature space dimensions are uncorrelated with respect to the variations accounted for. Previous experiments have suggested that this requirement is best met at the filter bank, since histogram normalization performs best at this signal analysis stage (cf. Section 7.1.4). Still the feature space might not only be distorted and translated, but also rotated by a small amount (e.g. Figure 7.1), which would not be treated properly by histogram normalization. In the following, a transformation will be proposed that is able to handle this type of mismatch between training and test data.

Just as in histogram normalization, training and test data of different conditions shall be transformed to some reference condition in order to reduce undesired variations in the speech signal. However, instead of mapping the axes of the feature space independently of each other, a linear transformation shall be applied to the complete acoustic vector. The aim is to reduce the differences between the condition-dependent covariance matrices in training and test.

To account for the type of mismatch depicted in Figure 7.1, the transformation will be restricted to be a rotation, which changes the orientation of the feature space axes but preserves Euclidean distances. The rotation will be further restricted to consist of elementary rotations that only map principal feature space axes.

First, a pathological case of an approximately “circular” feature space shall be considered where the reference and the condition-dependent covariance matrices are nearly diagonal with identical values. In this case, the eigenvectors will be oriented arbitrarily and the eigenvalues are all similar, which would result in undesired arbitrary rotations for different conditions. If, on the other hand, the feature space is elongated, i.e. if the scatter is non-uniform in different directions, at least some eigenvectors are well-defined. For this reason, the eigenvectors will be sorted in descending order of their eigenvalues, and a number of

elementary rotations will be applied. Only the first condition-dependent eigenvectors with dominantly larger eigenvalues will be mapped to their corresponding reference eigenvectors. Note that if all eigenvectors are considered at the same time, the transformation is identical to a principal component analysis (PCA), computed independently for training and test conditions.

7.2.2 Principle

Just as in histogram normalization, the reference condition has to be defined first. Here, the covariance matrix $\tilde{\Sigma}$ obtained from the full training corpus is used as reference. The corresponding D orthonormal reference eigenvectors $\tilde{v}_1, \ldots, \tilde{v}_D$ and eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_D$ are defined by:

$$\begin{aligned} \tilde{\Sigma}\tilde{v}_d &= \tilde{\lambda}_d \tilde{v}_d \qquad \tilde{v}_d \in \mathbb{R}^D, \quad ||\tilde{v}_d||^2 = 1, \quad d = 1, \ldots, D && (7.7) \\ \tilde{\lambda}_1 &\geq \tilde{\lambda}_2 \geq \ldots \geq \tilde{\lambda}_D \geq 0 && (7.8) \end{aligned}$$

The eigenvectors are sorted in descending order of their corresponding eigenvalues (Eqn. 7.8).

During normalization, the covariance matrix Σ_r of each training and test condition $r = 1, \ldots, R$ is computed from data X_r. Note that for improved readability the condition index r is omitted in all following equations.

The condition-dependent orthonormal eigenvectors and eigenvalues are calculated and sorted in the same way as the reference:

$$\begin{aligned} \Sigma v_d &= \lambda_d v_d \qquad v_d \in \mathbb{R}^D, \quad ||v_d||^2 = 1, \quad d = 1, \ldots, D && (7.9) \\ \lambda_1 &\geq \lambda_2 \geq \ldots \geq \lambda_D \geq 0 && (7.10) \end{aligned}$$

Note that the eigenvectors are unique except for a scale factor of ± 1. If the sign of each component of an eigenvector is inverted, the same eigenvector basis is obtained with one axis pointing into the opposite direction. Here the condition-dependent eigenvectors v_d are chosen such that the angles to the corresponding reference eigenvectors $\tilde{v}_d$ are less than or equal to 90 degrees, i.e. such that the dot product of all eigenvector pairs is positive:

$$\tilde{v}_d \cdot v_d \geq 0 \qquad d = 1, \ldots, D \tag{7.11}$$

A transformation matrix U_D that rotates all D condition-dependent eigenvectors to their corresponding reference eigenvectors is obtained by the product of the two eigenvector matrices $\tilde{V}$ and V (Eqn. 7.12). The first matrix is made of the reference eigenvectors $\tilde{v}_1, \ldots, \tilde{v}_D$, and the second is made of the condition-dependent eigenvectors $v_1, \ldots, v_D$:

$$\begin{aligned} U_D &= \tilde{V} \cdot V^T \qquad U_D, \tilde{V}, V \in \mathbb{R}^{D \times D} && (7.12) \\ \tilde{V} &= \left(\tilde{v}_1, \ldots, \tilde{v}_D\right) \\ V &= \left(v_1, \ldots, v_D\right) \end{aligned}$$

The matrix U_D is of little use, however, because it is to expect that only the direction of the first few eigenvectors is well-defined as described in the previous section. A transformation matrix that maps the first eigenvectors only will be constructed stepwise. First, the rotation matrix $\hat{U}_1$ to map the first condition-dependent eigenvector v_1 to the first reference eigenvector $\tilde{v}_1$ is derived. The rotation angle η_1 between the two eigenvectors is computed from their dot product, as they are of unit length:

$$\eta_1 = \arccos(\tilde{v}_1 \cdot v_1) \tag{7.13}$$

Since the two eigenvectors are not orthogonal, the Gram-Schmidt algorithm is applied to v_1 in order to obtain an orthonormal basis vector $\hat{v}_1$ lying in the same plane of rotation:

$$\hat{v}_1 = \frac{v_1 - (\tilde{v}_1 \cdot v_1) \cdot \tilde{v}_1}{||v_1 - (\tilde{v}_1 \cdot v_1) \cdot \tilde{v}_1||^2} \tag{7.14}$$

Next, the acoustic vector is projected onto the plane spanned by $\tilde{v}_1$ and $\hat{v}_1$ with the projection matrix J_1^T:

$$J_1^T = \left(\hat{v}_1, \tilde{v}_1\right)^T \qquad J_1^T \in \mathbb{R}^{D\times 2} \tag{7.15}$$

It is rotated within the plane with the rotation matrix R_1 (Eqn. 7.16) by the angle η_1, and projected back into the original $\mathbb{R}^{D\times D}$space with the transposed projection matrix $J_1 \in \mathbb{R}^{2\times D}$:

$$R_1 = \begin{pmatrix} \cos\eta_1 & \sin\eta_1 \\ -\sin\eta_1 & \cos\eta_1 \end{pmatrix} \qquad R_1 \in \mathbb{R}^{2\times 2} \tag{7.16}$$

Finally, a correction term $I - J_1 J_1^T$ with the identity matrix I has to be applied that restores the dimensions orthogonal to the plane of rotation lost in the first projection. It ensures that all these dimensions remain unchanged. The full rotation matrix $\hat{U}_1$ is derived by:

$$\hat{U}_1 = J_1 R_1 J_1^T + I - J_1 J_1^T \tag{7.17}$$

Since eigenvectors are orthogonal, it is possible to repeat the procedure sequentially for further eigenvector pairs. Each new transformation will have no impact on previous feature space rotations.

To compute the rotation matrix for the second pair of eigenvectors, the condition-dependent eigenvector v_2 is rotated by $\hat{U}_1$, and the corresponding orthonormal basis vector $\hat{v}_2$ is computed (Eqn. 7.18). Next the rotation angle η_2 (Eqn. 7.19) and the second rotation matrix $\hat{U}_2$ are derived (Eqn. 7.20). It rotates the feature space in the plane spanned by the second condition-dependent and the second reference eigenvector after the application of $\hat{U}_1$:

$$\hat{v}_2 = \frac{\hat{U}_1 v_2 - (\tilde{v}_2 \cdot \hat{U}_1 v_2) \cdot \tilde{v}_2}{||\hat{U}_1 v_2 - (\tilde{v}_2 \cdot \hat{U}_1 v_2) \cdot \tilde{v}_2||^2} \tag{7.18}$$

$$\eta_2 = \arccos(\tilde{v}_2 \cdot \hat{U}_1 v_2) \tag{7.19}$$

$$\hat{U}_2 = J_2 R_2 J_2^T + I - J_2 J_2^T \tag{7.20}$$

$$J_2 = \left(\hat{v}_2, \tilde{v}_2\right) \tag{7.21}$$

$$R_2 = \begin{pmatrix} \cos\eta_2 & \sin\eta_2 \\ -\sin\eta_2 & \cos\eta_2 \end{pmatrix} \tag{7.22}$$

The third and further eigenvectors can be mapped in the same way:

$$\hat{v}_d = \frac{U_{(d-1)} v_d - (\tilde{v}_d \cdot U_{(d-1)} v_d) \cdot \tilde{v}_d}{||U_{(d-1)} v_d - (\tilde{v}_d \cdot U_{(d-1)} v_d) \cdot \tilde{v}_d||^2} \tag{7.23}$$

$$\eta_d = \arccos(\tilde{v}_d \cdot U_{(d-1)} v_d) \tag{7.24}$$

$$\hat{U}_d = J_d R_d J_d^T + I - J_d J_d^T \tag{7.25}$$

$$J_d = \left(\hat{v}_d, \tilde{v}_d\right) \tag{7.26}$$

$$R_d = \begin{pmatrix} \cos\eta_d & \sin\eta_d \\ -\sin\eta_d & \cos\eta_d \end{pmatrix} \tag{7.27}$$

$$U_d = \hat{U}_d \hat{U}_{(d-1)} \hat{U}_{(d-2)} \ldots \hat{U}_1 \tag{7.28}$$

The product of all rotation matrices $\hat{U}_1, \ldots, \hat{U}_d$ (Eqn.7.28) yields the condition-dependent transformation matrix U_d that maps the first d eigenvectors. U_d is equivalent to the parameter set α_r (Eqn. 2.9) of the transformation function $f_\alpha(x)$ (Eqn. 2.11) defined in Section 2.3. It is applied to normalize the acoustic vectors by:

$$\begin{aligned} x \to \tilde{x} &= f_\alpha(x) \\ &= U_d \cdot x \end{aligned} \tag{7.29}$$

In the D-dimensional feature space, up to $D-1$ rotations may be carried out. The last dimension matches automatically due to the orthogonality constraint, which is a nice consistency check for the procedure. The deviation angle η_D between the rotated Dth condition-dependent eigenvector and the Dth reference eigenvector needs to be zero, and the resulting rotation matrix $U_{(D-1)}$ must be identical to the matrix U_D derived by equation Eqn. 7.12.

7.2.3 Experimental Results

Feature space rotation may be applied at the same signal analysis stages as histogram normalization (cf. Section 7.1.4). If applied at the filter bank it now makes a difference whether the feature space is normalized before or after log compression. Rotation before log compression may result in negative coefficients, which prevents the successive application of the logarithm. For this reason, rotation was only applied after log compression.

To calculate the reference condition, the covariance matrix $\tilde{\Sigma}$ and the eigenvector basis $\tilde{v}_1, \ldots, \tilde{v}_D$ are computed on the full training corpus (cf. Eqn. 7.7). It turns out that at the log filter bank stage the first eigenvalue is significantly larger than all others as shown in Figure 7.8 for the VerbMobil II training corpus. Note that the logarithm *increases* the scatter of the filter bank coefficients, as these are typically small. The feature space has apparently one preferred direction with large scatter, and along the other principal axes data scatter is much smaller [Molau & Hilger$^+$ 02]. For this reason, recognition tests where only the first condition-dependent eigenvector v_1 is mapped to the first reference eigenvector $\tilde{v}_1$ are carried out first.

During normalization, the covariance matrix Σ and the eigenvector basis $v_1, \ldots, v_D$ are calculated for each training and test condition $r = 1, \ldots, R$ (cf. Section 7.1.2). The first condition-dependent eigenvector v_1 deviates typically by a few degrees from the direction of the first reference eigenvector $\tilde{v}_1$ as shown in Figure 7.9. The figure depicts a histogram over the condition-wise deviation angles η_1 (cf. Eqn. 7.13) calculated on log filter bank vectors of the VerbMobil II training corpus.

Finally, the condition-dependent rotation matrix U_1 for the first eigenvector is derived (cf. Eqn. 7.17) and the training and test data are transformed (cf. Eqn. 7.29). A normalized acoustic model is trained (cf. Eqn. 2.10), and the normalized test vectors are used in recognition.

Recognition test results for feature space rotation at different signal analysis stages are summarized in Table 7.6. Given are the word error rate, the mean ratio of the first two reference eigenvalues $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$, and the mean deviation angle $\overline{\eta_1}$ between the first condition-dependent eigenvectors v_1 and the first reference eigenvector $\tilde{v}_1$.

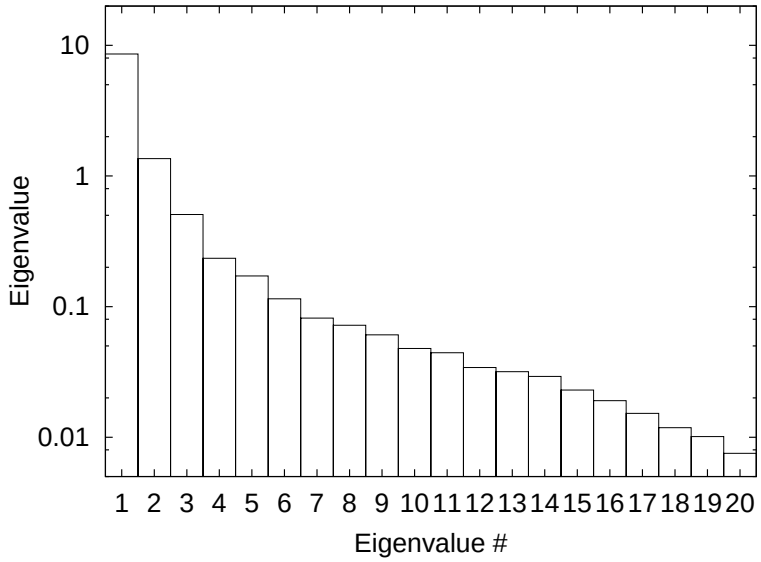

Figure 7.8: Sorted eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_D$ of the reference covariance matrix $\tilde{\Sigma}$ computed on log filter bank coefficients of the VerbMobil II training corpus. Note the logarithmic scale of the ordinate.

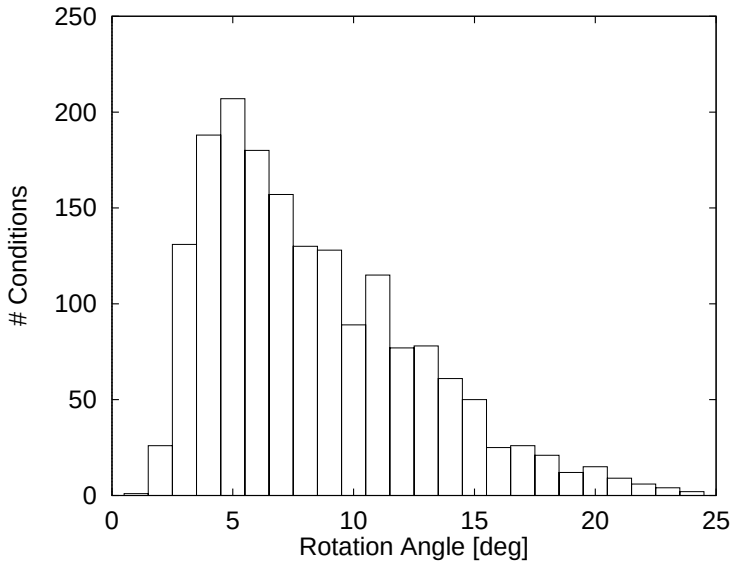

Figure 7.9: Histogram over the deviation angles η_1 between the first eigenvectors v_1 of the condition-dependent covariance matrices and the first reference eigenvector $\tilde{v}_1$ computed on log filter bank coefficients of the VerbMobil II training corpus.

Rotation of log filter bank vectors yielded a clear improvement in recognition accuracy. The gain was as large as for basic histogram normalization, but smaller than for histogram normalization with silence fraction treatment.

At the cepstrum stage, the ratio of the first two reference eigenvalues was much smaller. Consequently, the principal axes were not as well-defined, the deviation angles increased significantly, and the recognition performance dropped. A detailed analysis of recognition errors revealed that the performance improved for most conditions with small rotation angles, but the word error rate almost doubled for a few test conditions with rotation angles η_1 close to 90 degrees, indicating a change in the order of eigenvectors.

When rotating the feature space after linear discriminant analysis, the mean ratio of the first and second reference eigenvalues and the average deviation angle of the first eigenvectors were similar to the values observed for the filter bank stage. The reference

Table 7.6: Recognition test results on the VerbMobil II DEV99B corpus for feature space rotation at different signal analysis stages to map the first eigenvector. Given are the mean ratio of the first two reference eigenvalues, the mean deviation angle between the first condition-dependent and the first reference eigenvector, and the word error rate.

Normalization Stage	Mean Eigenvalue Ratio $\tilde{\lambda}_1/\tilde{\lambda}_2$	Mean Deviation Angle $\overline{\eta_1}$ [deg]	Overall [%] Del - Ins	WER
baseline without normalization			4.9 - 4.4	24.6
log filter bank	6.3	8.4	4.6 - 4.3	**23.0**
cepstrum	1.4	32.9	5.3 - 5.0	27.8
after LDA	5.4	8.6	5.1 - 4.2	24.1

covariance matrix $\tilde{\Sigma}$ was diagonal and the corresponding eigenvector matrix $\tilde{V}$ was the identity matrix, which results from the property of linear discriminant analysis to uncorrelate the feature space dimensions. Normalization after LDA gave the same minor performance improvement than histogram normalization at this stage (cf. Table 7.2), but it was inferior to rotation at the log filter bank stage. Hence, the outcome was comparable to histogram normalization which performed best at the log filter bank stage as well. Further tests were carried out at this signal analysis stage only.

Note that the order of mean normalization and feature space rotation does not matter as long as both are applied condition-wise. Even though the mean vector will be different, the resulting acoustic vector is identical regardless whether first the mean is subtracted from the vector before it is rotated, or whether mean subtraction is carried out after rotation. In practice, mean (and possibly variance) normalization are applied sentence-wise or in a sliding window. Informal tests on different corpora have shown, however, that similar to histogram normalization (cf. Section 7.1.7) cepstral mean/variance normalization after rotation was typically slightly superior to mean normalization before rotation.

In a next set of experiments, the number of condition-dependent eigenvectors mapped to their corresponding reference eigenvectors was increased. Matching more than the first eigenvector further reduces the mismatch between the condition-dependent covariance matrices Σ and the reference covariance matrix $\tilde{\Sigma}$. However, limits are set by the discrete order of eigenvectors. The smaller the differences between subsequent eigenvalues, the larger is the chance that the order of eigenvectors changes and for some conditions principal axes are mapped that represent different acoustic characteristics.

In practice, it turned out that even the direction of the second principal axis is not well defined, and that the order of eigenvectors changes for different conditions. The rotation angles η_d for the second and further pairs of eigenvectors increase significantly as shown for the VerbMobil II corpus in Figure 7.10. On other corpora they soon became as large as 90 degree, and a rotation was not sensible.

The corresponding recognition test results are summarized in Table 7.7. They show that mapping more than the first eigenvector does not improve the recognition accuracy. In

Table 7.7: Recognition test results on the VerbMobil II DEV99B corpus for feature space rotations to map up to four eigenvectors. The given mean deviation angles refer to the highest mapped eigenvector.

Feature Space Rotation	Mean Deviation Angle $\overline{\eta_d}$ [deg]	Overall [%] Del - Ins	WER
baseline without normalization		4.9 - 4.4	24.6
first eigenvector	8.4	4.6 - 4.3	**23.0**
first two eigenvectors	10.7	4.6 - 4.3	23.2
first three eigenvectors	19.8	4.0 - 4.6	23.0
first four eigenvectors	21.6	4.3 - 4.8	23.6

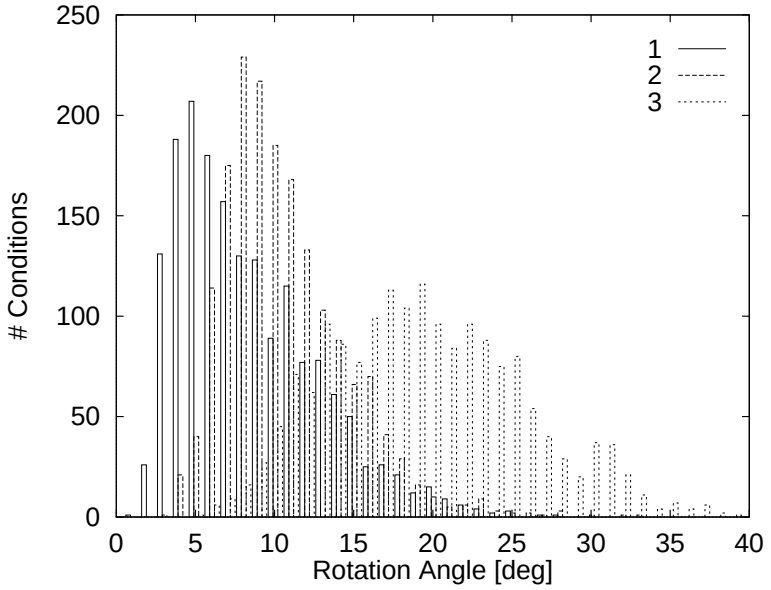

Figure 7.10: Histogram over the deviation angles $\eta_1, \ldots, \eta_3$ between the first three condition-dependent eigenvectors $v_1, \ldots, v_3$ and the first reference eigenvectors $\tilde{v}_1, \ldots, \tilde{v}_3$ computed on log filter bank coefficients of the VerbMobil II training corpus.

fact, whereas on the VerbMobil II corpus the word error rate was of the same order when the first two or three eigenvectors were mapped, the performance significantly deteriorated in these cases on other corpora.

7.3 Combination of Histogram Normalization and Rotation

7.3.1 Motivation

Since feature space rotation overcomes one of the principal limits of histogram normalization, it is interesting to see if the gain in recognition performance obtained by applying both techniques is additive.

The natural order of normalization would be to rotate the acoustic vectors first for optimal orientation of the feature space dimensions, and then normalize the distribution of each dimension. On the other hand, histogram normalization with silence fraction treatment gives a larger gain in recognition performance than feature space rotation. The deviation angles η_1 between the first condition-dependent eigenvectors v_1 and the first reference eigenvector $\tilde{v}_1$ are significantly reduced when histogram normalization is applied before rotation (Figure 7.11), which could make the estimation of the rotation plane and angle more reliable and give superior results. From this perspective, feature space rotation would be concerned with mismatch in the condition-dependent distributions that remains after histogram normalization.

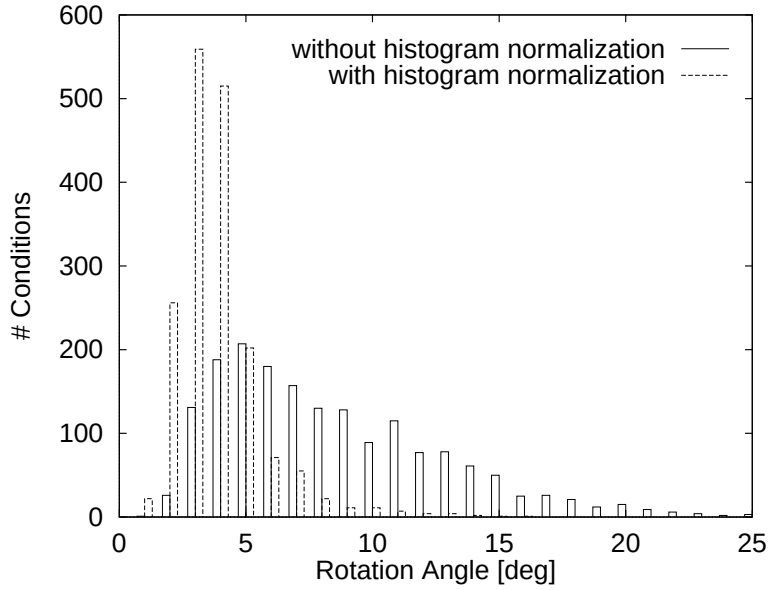

Figure 7.11: Histogram over the deviation angles η_1 between the first condition-dependent eigenvectors v_1 and the first reference eigenvector $\tilde{v}_1$ estimated on log filter bank vectors of the VerbMobil II training corpus. Results are given both with and without histogram normalization before rotation.

7.3.2 Experimental Results

Recognition test results with either normalization technique in different order are summarized in Table 7.8. Better results were achieved when histogram normalization was applied before feature space rotation. The word error rate was lower than for feature space rotation alone (23.0%), but in the case of VerbMobil II not as low as for histogram normalization with silence fraction treatment (21.8%). Both techniques seem to account for the same speech signal variations (which can be seen by the reduced rotation angles, Figure 7.11), but histogram normalization is more efficient.

Table 7.8: Recognition test results on the VerbMobil II DEV99B corpus for the combination of feature space rotation to map the first eigenvector and histogram normalization with silence fraction treatment.

Normalization		Overall [%]	
First Stage	Second Stage	Del - Ins	WER
baseline without normalization		4.9 - 4.4	24.6
rotation	histogram normalization	4.6 - 4.1	22.8
histogram normalization	rotation	4.6 - 3.8	**22.4**

7.4 Normalization under Different Mismatch Conditions

Histogram normalization with silence fraction treatment, feature space rotation to map the first eigenvector, and a combination of both techniques at the log filter bank stage has been evaluated on different corpora to analyze the performance of these techniques under various degrees of mismatch between training and test data.

The VerbMobil II corpus contains a minor acoustic mismatch as one part of the training data was collected with a close-talking, the other with a room microphone. In addition, the test data were from a different domain and recorded with a different microphone than most of the training data (cf. Section 5.2.2). Recognition test results were presented in Tables 7.4, 7.6 and 7.8. Histogram normalization with silence fraction treatment reduced the word error rate by 11% relative and feature space rotation by 7% relative. A combination of both techniques gave no further gain in recognition performance.

Recognition tests with the best normalization setup were repeated for an across-word system. The results are presented in Table 7.9. Even though histogram normalization with silence fraction treatment still gave a significant improvement in recognition accuracy, it was smaller than the gain in the case of within-word modeling similar to the combination of vocal tract length normalization and across-word models (cf. Section 6.6).

Recognition test results for the EuTrans II corpus are summarized in Table 7.10. Both the training and test data were recorded over wireline telephone, so that there is no explicit acoustic mismatch (cf. Section 5.3.1). However, the channel quality varied significantly between different recording sessions. Basic histogram normalization without silence fraction treatment gave no improvement in recognition accuracy on this corpus. The transmission channel was more noisy and showed larger variations from one condition to the next, which is why deviations from the average silence fraction may have had a larger impact on the recognition accuracy. Silence fraction adapted histogram normalization yielded a relative error rate reduction of 5%, and feature space rotation improved the recognition performance by a similar amount. The reductions of both techniques were again not additive.

Table 7.9: Across-word system recognition test results on the VerbMobil II DEV99B corpus. Given are word error rates for the optimized baseline system without normalization, and for histogram normalization with silence fraction treatment.

Normalization	Overall [%]	
	Del - Ins	WER
baseline without normalization	4.6 - 3.8	21.6
histogram normalization	4.3 - 3.3	**20.2**

Table 7.10: Recognition test results on the EuTrans II corpus for histogram normalization with silence fraction treatment, feature space rotation to map the first eigenvector, and a combination of both techniques.

Normalization		Overall [%]	
First Stage	Second Stage	Del - Ins	WER
baseline without normalization		4.2 - 3.1	16.5
histogram normalization	-	3.8 - 3.0	15.6
rotation	-	3.6 - 3.1	15.8
rotation	histogram normalization	3.7 - 3.1	**15.5**
histogram normalization	rotation	3.5 - 3.1	15.6

The CarNavigation database is a task with large mismatch conditions. The training data were recorded in a quiet office environment, and two of the test sets were recorded in cars (city and highway traffic, cf. Section 5.3.2). In scenarios with such a mismatch there is much room for improvements. A standard normalization technique is cepstral variance normalization. On this task, it lowered the recognition accuracy in the clean office condition, but clearly improved the baseline result for the city and highway test sets (Table 7.11).

Histogram normalization reduced the word error rate significantly both with and without variance normalization. Better results were obtained without this extra normalization step. The variance of the filter bank channels is already implicitly normalized when the feature space dimensions are mapped onto the same reference histogram, which is why

Table 7.11: Recognition test results on the CarNavigation test corpora for histogram normalization with silence fraction treatment, feature space rotation to map the first eigenvector, and a combination of both techniques. Results are reported with and without subsequent cepstral variance normalization (CVN).

Normalization			WER [%]		
CVN	First Stage	Second Stage	Office	City	Highway
yes	baseline without normalization		4.2	20.8	39.7
	histogram	-	3.6	12.4	21.4
	rotation	-	3.8	11.7	21.0
	rotation	histogram	3.7	10.4	19.0
	histogram	rotation	**3.5**	**8.9**	**14.6**
no	baseline without normalization		2.9	31.6	74.2
	histogram	-	2.6	8.2	14.3
	rotation	-	2.5	24.0	64.6
	rotation	histogram	**2.4**	9.5	18.0
	histogram	rotation	2.9	**7.1**	**11.1**

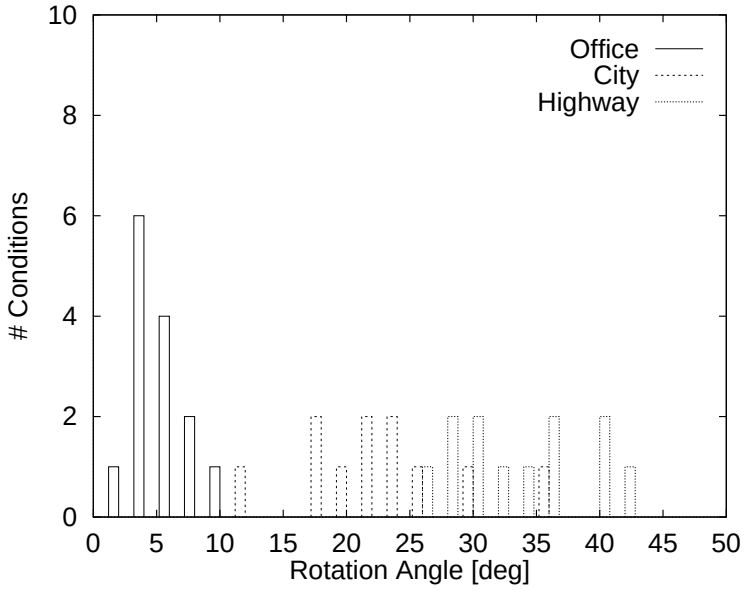

Figure 7.12: Histogram over the deviation angles η_1 between the first eigenvectors v_1 of the condition-dependent covariance matrices and the first reference eigenvector $\tilde{v}_1$ computed on log filter bank coefficients of the different CarNavigation test corpora. The rotation angles increase with the mismatch between training and test data.

a further transformation to unity cepstral variance may be counterproductive. Without variance normalization, the word error rate was reduced between 10% relative (office) and 81% relative (highway).

When feature space rotation was applied, the rotation angles for the test data increased with the mismatch. Whereas on the office data the mean rotation angle $\overline{\eta_1}$ was 6 degrees, it increased to 23 degrees on the city and 32 degrees on the highway data (Figure 7.12).

In connection with cepstral variance normalization, feature space rotation even outperformed histogram normalization slightly. The reduction in word error rate varied between 10% relative (office) and 47% relative (highway). Without variance normalization, however, rotation gave only small improvements over the baseline system. This supports the notion that even though similar variations are accounted for by histogram normalization and feature space rotation, the mismatch is reduced in a different way. In particular the variance of the acoustic signal cannot be handled properly by feature space rotation alone. This comes as no surprise, as the feature space axes are rotated but not scaled.

If applied in the right order, feature space rotation and histogram normalization together performed better than both techniques alone. The best result on the office test data was achieved when rotation was applied before histogram normalization, and on the city and highway data when applied afterwards. The experiments show that the normalization method that gives most gain in recognition performance should be applied first.

7.5 Summary

Histogram normalization and feature space rotation are normalization techniques in the acoustic feature space. They aim at reducing the mismatch between training and test data by mapping different conditions (different speakers, speaking styles, transmission channels, etc.) to some reference condition. They are model-free and text-independent, i.e. they only rely on global statistics of the speech signal.

It was shown that histogram normalization is conceptually simple but improves the recognition accuracy on a variety of corpora. It can be applied to different signal analysis stages and performs best when log filter bank vectors are transformed. Normalization of training and test data was superior to normalization of test data alone. The larger the acoustic mismatch between the recording conditions in training and test, the larger was the gain in recognition performance. This suggests that histogram normalization can reduce channel and environmental variations very efficiently.

Feature space rotation is a normalization technique proposed to relax the assumption of histogram normalization regarding the orientation of the feature space axes. It aims at reducing the mismatch between condition-dependent covariance matrices and a reference covariance matrix. For this purpose, transformation matrices were derived that map the principal axes with the largest data scatter.

It was shown that at different signal analysis stages the feature space is not uniform but has one preferred direction with especially large scatter. Feature space rotation to match the first eigenvector performed best at the log filter bank stage similar to histogram normalization. Matching subsequent principal axes with large scatter did not improve the recognition accuracy any further. On the three corpora under investigation, the reduction in word error rate by feature space rotation was typically somewhat lower than the reduction by histogram normalization with silence fraction treatment.

In the case of major mismatch between training and test data, further improvements of recognition accuracy were achieved by a combination of histogram normalization and feature space rotation. Best results were obtained when first the normalization method was applied that performs alone better.

Chapter 8

Combination of Normalization Schemes

8.1 Motivation

Vocal tract length normalization and histogram normalization/feature space rotation are all carried out during signal analysis and interact in a complex way. VTN frequency axis warping is applied to the magnitude spectrum, i.e. before the filter bank and therefore before other normalization schemes. However, the warping factor estimation relies on likelihood calculations based on the final acoustic vector, i.e. after all other normalization schemes. Hence, a joint optimization of the warping factor and the speaker-dependent histograms/covariance matrices would require a complex iterative optimization.

In Chapter 6 it was shown that vocal tract length normalization has only a single free parameter which can be estimated reliably on small data samples. Histogram normalization and feature space rotation are based on a reference histogram and covariance matrix, which were obtained somewhat arbitrary from the overall training data distribution as discussed in Section 7.1.3. For these reasons it makes sense to estimate the warping factors and transformation functions independently of each other on unnormalized data. A number of informal recognition tests on the VerbMobil II corpus have confirmed that this variant performs better than estimating warping factors on histogram-normalized data, or computing speaker-dependent histograms and covariance matrices on VTN-normalized data.

8.2 Experimental Results for Different Corpora

Recognition test results for VerbMobil II are summarized in Table 8.1. Vocal tract length normalization was applied in combination with histogram normalization with silence fraction treatment, since additional feature space rotation was not helpful on this corpus. The word error rate could be reduced by a small amount to the overall best result of 21.6%, i.e. the gain in recognition performance was only to a small extent additive. The reduction in word error rate was 12% relative to the baseline system.

Table 8.1: Recognition test results on the VerbMobil II DEV99B corpus for a combination of vocal tract length normalization and histogram normalization with silence fraction treatment.

Normalization		Overall [%]	
VTN	Histogram	Del - Ins	WER
baseline without normalization		4.9 - 4.4	24.6
fast	no	4.4 - 4.7	22.7
two-pass	no	4.6 - 4.6	22.6
fast	yes	4.8 - 4.1	**21.6**
two-pass	yes	4.3 - 4.4	21.7

In case of EuTrans II, fast and two-pass vocal tract length normalization reduced the word error rate by 8% relative, and histogram normalization with silence fraction treatment by 5% relative (cf. Table 7.10). Again, a combination of VTN and histogram normalization was tested, as feature space rotation did not help in connection with histogram normalization. On this corpus, the gain in recognition performance was to a large extent additive and yielded again an overall reduction of up to 12% relative to the baseline system (Table 8.2).

Table 8.3 summarizes recognition test results for different normalization schemes on the CarNavigation corpus. Only results without variance normalization are given here, because they were better than the corresponding results with variance normalization except for the baseline setup.

Two-pass vocal tract length normalization with speaker-wise warping factor estimation reduced the word error rate by 21% relative on the clean office data. Similar large reductions for "simple" tasks (here isolated-word recognition in clean office environments) were reported for VTN by other groups as well (cf. Section 3.1.1). On the city condition the gain reduced to 13%, and on the highway data there was essentially no improvement in recognition accuracy at all. The main reason is that warping factors cannot be estimated reliably when the word error rate in the first recognition pass is well above 50%.

Table 8.2: Recognition test results on the EuTrans II corpus for a combination of vocal tract length normalization and histogram normalization with silence fraction treatment.

Normalization		Overall [%]	
VTN	Histogram	Del - Ins	WER
baseline without normalization		4.2 - 3.1	16.5
fast	no	3.9 - 3.0	15.4
two-pass	no	3.7 - 3.3	15.1
fast	yes	3.5 - 2.9	**14.5**
two-pass	yes	3.5 - 2.8	14.8

Table 8.3: Recognition test results on the CarNavigation corpora for a combination of vocal tract length normalization, histogram normalization with silence fraction treatment, and feature space rotation to map the first eigenvector. Results are reported without subsequent cepstral variance normalization.

Normalization			WER [%]		
VTN	First Stage	Second Stage	Office	City	Highway
baseline without normalization			2.9	31.6	74.2
fast	-	-	2.8	27.5	67.4
two-pass	-	-	2.3	27.5	73.1
fast	rotation	histogram	4.1	11.2	17.7
two-pass			2.9	10.4	16.7
fast	histogram	rotation	2.8	6.8	11.1
two-pass			**2.2**	**6.6**	**10.4**

Fast VTN, on the contrary, is text-independent and does therefore not depend on the word error rate. Warping factor estimation is based on Gaussian mixture models that describe the distribution of acoustic vectors in the feature space (cf. Section 6.3.3). It is surprising that the warping factor can still be estimated reliably when the distribution of the training data differs significantly from the test data distribution (mismatch in the city and highway condition). In connection with variance normalization, fast VTN was superior to two-pass VTN on all conditions. Without variance normalization, the performance was typically somewhat worse. The explanation for slightly more robust warping factor estimates in the first case might be that variance normalization reduces the mismatch between the Gaussian mixture models trained on office data and noisy acoustic vectors in test.

The gain in recognition accuracy by vocal tract length normalization and histogram normalization/feature space rotation was to some extent additive on the CarNavigation corpus. This is consistent with the underlying model that the former technique accounts for speaker-dependent variations only, whereas the latter normalization schemes account for speaker and environmental variations. The overall best results were achieved by two-pass VTN followed by histogram normalization and feature space rotation.

8.3 Summary

Normalization is a powerful technique to increase the robustness of automatic speech recognition systems. Irrelevant variations caused by varying transducers and transmission channels, speakers and speaking styles, as well as ambient or channel noise are reduced or completely removed.

In scenarios with a large acoustic mismatch between training and test data the effects are especially large. As demonstrated in Table 8.4 for the isolated-word CarNavigation corpus, normalization can make the difference from essentially zero recognition accuracy to an acceptable level where 90% of all words are correctly recognized. In conditions with only a speaker mismatch the gain in recognition performance is lower, but the word error still decreases significantly.

Table 8.4: Effects of different normalization steps on the CarNavigation test corpora. Results are given for cepstral mean normalization (CMN), histogram normalization (HN) with silence fraction treatment (HNSIL), feature space rotation to map the first eigenvector (ROT), and two-pass vocal tract length normalization (VTN).

Normalization Steps	WER [%]		
	Office	City	Highway
baseline without normalization	2.8	68.0	99.0
CMN	2.9	31.6	74.2
CMN, HN	2.8	10.2	16.6
CMN, HNSIL	2.6	8.2	14.3
CMN, HNSIL, ROT	2.4	7.1	11.1
CMN, HNSIL, ROT, VTN	**2.2**	**6.6**	**10.4**

Chapter 9

Scientific Contributions

The aim of this work was to develop and improve normalization techniques in the acoustic feature space to remove undesired variations from the acoustic signal and increase the performance of automatic speech recognition systems.

A classification scheme for different normalization and adaptation schemes was introduced in this work. Based on a model for training and test, common properties of normalization and adaptation as well as differences between these techniques could be explained. Adaptive acoustic modeling was introduced into the mathematical framework of statistical speech recognition, and an overview of normalization techniques proposed in the literature was given.

Vocal tract length normalization was one normalization technique studied in detail.

The first goal was to achieve a *consistently large gain in recognition performance under variable environments*. A number of optimizations and improvements of the baseline two-pass VTN approach were implemented and tested. Helpful was the weighting of acoustic vectors with their energy during warping factor estimation, and the re-estimation of the phonetic decision tree and the LDA transformation matrix on normalized training data. A vocal tract length normalization scheme in training and test was developed that proved to yield consistently good performance on all corpora under investigation:

- on the VerbMobil II corpus, a 10k-word vocabulary German conversational speech task with many spontaneous speech phenomena, the word error rate was reduced from 25.7% to 23.3%
- on the 20k-word vocabulary North American Business News corpus which is made of read English texts recorded in clean acoustic conditions, the word error rate dropped from 12.5% to 11.6%
- on the 2k-word vocabulary EuTrans II corpus of Italian spontaneous speech recorded over a telephone channel of variable quality, the word error rate was reduced from 16.5% to 15.1%

In summary, consistent improvements of 8% to 9% relative were achieved by two-pass recognition on a large variety of conditions, namely different vocabulary sizes, languages,

speaking styles, and recording environments. The word error rate achieved on the aforementioned corpora was similar to the results obtained by across-word modeling (23.3%, 11.5% and 15.7%, respectively), but the gain of both techniques was not fully additive.

In this work it was shown that the *full gain in recognition accuracy* by vocal tract length normalization can be obtained *without an increase in computation time*. Two different Gaussian mixture model based techniques which are text-independent and require only a single recognition pass were studied. Warping factor estimation employing GMMs with pooled variance vector trained on unnormalized data yielded the desired performance. On all three corpora, the word error rate of text-independent fast VTN was virtually identical to the two-pass recognition results:

- 23.5% for fast compared to 23.3% for two-pass VTN on the VerbMobil II corpus
- 11.5% for fast compared to 11.6% for two-pass VTN on the North American Business News corpus
- 15.3% for fast compared to 15.1% for two-pass VTN on the EuTrans II corpus

At identical pruning settings, fast VTN demanded even less computation time than the baseline system without normalization as the overhead from the warping factor estimation was more than compensated by more efficient pruning with normalized acoustic vectors and models.

The requirements of vocal tract length normalization for *online recognition* were met by incremental warping factor estimation. Even though this special type of fast VTN comes at the cost of a slightly reduced gain in recognition performance, it does not introduce any additional delay in signal analysis which is necessary for online recognition systems. Fast VTN with incremental warping factor estimation was successfully applied in the RWTH speech recognition system used in the final VerbMobil II evaluation.

A novel *integrated frequency axis warping approach* was developed that merges a number of successive signal analysis steps into a single one. The filter bank can be omitted, the logarithm is applied directly to the spectral lines, and all frequency axis warping schemes such as Mel-frequency warping and vocal tract length normalization are integrated into the cepstrum transformation. The approach avoids possible quantization and interpolation problems of other techniques and yields a compact implementation of Mel-frequency cepstral coefficients by a simple matrix multiplication of the log-magnitude spectrum. On the VerbMobil II and North American Business News corpora, integrated frequency axis warping yielded the same recognition performance as the traditional approach with filter bank.

It was shown that integrated frequency axis warping allows for a *better control over the amount of spectral smoothing* than a fixed filter bank. Increasing the number of cepstral coefficients without enlarging the acoustic vector did not change the recognition accuracy on the North American Business News corpus, but lowered the word error rate on the

VerbMobil II corpus from 25.7% to 24.9%.

Histogram normalization and *feature space rotation* were the second focal point of this work. They aim at reducing the mismatch between training and test data by mapping different conditions (different speakers, speaking styles, transmission channels, etc.) to some reference condition. Both techniques are model-free and text-independent, i.e. they only rely on global statistics of the speech signal.

Histogram normalization is widely used in image processing, but the application in automatic speech recognition has been largely unexplored. It aims at reducing the mismatch by mapping the condition-dependent cumulative distributions of each feature vector component to some reference distribution. In this work it was shown that histogram normalization is conceptually simple but improves the recognition accuracy on a variety of corpora. It was applied to *different signal analysis stages*, namely to the filter bank, to Mel-frequency cepstral coefficients, and to the LDA-transformed acoustic vector. Sequential normalization at different signal analysis stages was studied as well. It turned out that histogram normalization performed best when log filter bank vectors were transformed. *Normalization of training and test data* was superior to normalization of the test data alone. On the VerbMobil II corpus, the word error rate could be reduced from 24.6% to 23.0% by basic histogram normalization in training and test.

Smoothing of the reference histogram was successful in the case of basic histogram normalization and reduced the word error rate further to 22.5% on the VerbMobil II corpus. Even though sequential normalization at different signal analysis stages was in most cases to some extent additive, it could not reduce the word error rate below this value.

Histogram normalization relies on the assumption that global statistics of the acoustic signal are identical independently of what is actually spoken. This requirement was relaxed by the new approach of explicit *silence fraction treatment*. It was shown that the recognition accuracy is significantly improved if the reference histogram is adapted to the silence fraction of each condition. Histogram normalization with silence fraction treatment reduced the word error rate also on corpora with only a minor or no mismatch between training and test data. The gain in recognition accuracy depended on the amount of mismatch:

- on the VerbMobil II corpus with a minor acoustic and scenario mismatch, the word error rate reduced from 24.6% to 21.8%
- on the EuTrans II corpus with no mismatch but variable telephone channel conditions, the word rate rate dropped from 16.5% to 15.6%
- on the CarNavigation corpus with a large mismatch (training data collected in quiet office environment, test data recorded in cars), the word error rate reduced from 2.9% to 2.6% on the office test set, from 31.6% to 8.2% on the city test set, and from 74.2% to 14.3% on the highway test set

Feature space rotation is a normalization technique proposed to relax the assumption of histogram normalization regarding the orientation of the feature space axes. It aims at reducing the mismatch between condition-dependent covariance matrices and a reference covariance matrix. For this purpose, transformation matrices were derived that map the principal axes with the largest data scatter.

It was shown in this work that the feature space is not uniform at different signal analysis stages, but has one preferred direction with especially large scatter. Feature space rotation to match the first eigenvector performed best at the log filter bank stage similar to histogram normalization. Matching subsequent principal axes with large scatter did not improve the recognition accuracy any further. On the three corpora under investigation, the reduction in word error rate by feature space rotation alone was typically somewhat lower than the reduction by histogram normalization with silence fraction treatment:

- on the VerbMobil II corpus, the word error rate reduced from 24.6% to 23.0%
- on the EuTrans II corpus, the word rate rate dropped from 16.5% to 15.8%
- on the CarNavigation corpus, the word error rate reduced from 2.9% to 2.5% on the office test set, from 31.6% to 24.0% on the city test set, and from 74.2% to 64.6% on the highway test set

A *combination of histogram normalization and feature space rotation* was helpful in the case of large acoustic mismatch (CarNavigation). On the other two corpora, the combination performed at best as good as histogram normalization with silence fraction treatment alone. Both techniques seem to account for the same speech signal variations, but histogram normalization is more efficient. It was found that in general the normalization technique that performs best on its own should be applied first, i.e. typically histogram normalization before feature space rotation.

The gain of vocal tract length normalization and histogram normalization/feature space rotation was to some extent additive. This supports the notion that VTN removes speaker-specific variations, whereas histogram normalization and feature space rotation account for both speaker- and environment dependent variations. The *combination of the best setups* developed in this work yielded the following improvements in recognition accuracy:

- on the VerbMobil II corpus, the word error rate could be reduced by 12% relative from 24.6% to 21.6% with fast VTN and histogram normalization with silence fraction treatment
- on the EuTrans II corpus, the reduction in word rate rate was as well 12% relative from 16.5% to 14.5% with fast VTN and histogram normalization with silence fraction treatment
- on the CarNavigation corpus, the word rate reductions were 24% relative on the office test set (2.9% to 2.2%), 79% relative on the city test set (31.6% to 6.6%) and 86% relative on the highway test set (74.2% to 10.4%). The reductions were achieved by a combination of two-pass VTN, histogram normalization with silence fraction treatment, and feature space rotation.

Chapter 10

Outlook

Different normalization techniques were studied in this work. Whereas vocal tract length normalization was developed to a point were consistently large improvements are obtained without an increase of computation time, there are still a number of open questions with respect to integrated frequency axis warping. It was shown that increasing the number of Mel-frequency cepstral coefficients helps to improve the recognition accuracy on the VerbMobil II corpus, but not on the North American Business News corpus. A more detailed investigation may reveal which information relevant to the recognition process is captured in higher cepstrum coefficients.

It might also be interesting to combine the approach with the linear transformation based signal analysis proposed by Yu and Waibel [Yu & Waibel 00]. It has been shown by different authors that the Mel-scale increases the recognition accuracy, so integrated Mel-frequency warping may improve an otherwise fully data-driven signal analysis front-end.

The integrated frequency axis warping approach may help to derive an analytic expression for proper handling of the Jacobian determinant in vocal tract length normalization that is so far omitted during warping factors estimation in training. First experiments in that direction have been reported in [Pitz & Molau$^+$ 01]. Taking the Jacobian determinant into account may improve the warping factor estimation and yield superior recognition performance.

Histogram normalization as proposed in this work requires several sentences per speaker to estimate the condition-dependent distributions reliably. Informal test have shown that single sentences are not sufficient. To reduce the amount of data required for histogram estimation, either a parametric transformation functions with larger bin sizes [Hilger & Ney 01] or efficient smoothing techniques may be helpful. So far, smoothing of the reference histogram was successful at intermediate stages only, and smoothing of the condition-dependent histograms estimated on sparse data was not helpful at all. More elaborate smoothing techniques may combine the robustness of parametric approaches on little adaptation data with the advantage of being not restricted to a specific transformation function.

In this work, the histogram over all training data is taken as the reference histogram to which all training and test data are mapped. It cannot be ruled out, however, that other distributions are more appropriate for reference. The same holds for the reference covariance matrix which is used to define the direction of the principle axes for feature space rotation.

Text-dependent transformation may be a way to improve the recognition gain by histogram normalization beyond the silence fraction treatment proposed in this work. Phoneme-dependent transformation may give a larger degree of freedom and help to remove further speaker-dependent variations. A similar approach was investigated in [Padmanabhan & Dharanipragada 01], but there only the test data were normalized at the cepstrum stage.

It was shown that taking the silence fraction into account improves the basic histogram normalization approach. The same technique may be applied in feature space rotation: Two reference covariance matrices may be derived for speech and silence frames, and a silence-fraction adapted reference covariance matrix could be obtained by linear interpolation between these. In addition, the current approach of mapping the principal axes with the largest data scatter could be extended by an additional scaling of the axes to match their eigenvalues as well.

The complex interaction between vocal tract length normalization and histogram normalization/feature space rotation, in particular a joint optimization of the warping factor, histogram and covariance matrix, may be studied in greater detail. Informal tests have shown that the independent parameter estimation pursued in this work is a good choice, but an iterative optimization may still be superior.

Finally, a more detailed comparison of the performance of normalization techniques studied here with adaptation techniques would be desirable. Especially for the CarNavigation task it would be interesting to see how maximum likelihood linear regression or similar adaptation techniques perform. A combination of normalization techniques with adaptation of the acoustic model may be another way to improve the recognition accuracy.

Bibliography

[Acero & Stern 91] A. Acero, R. M. Stern: Robust Speech Recognition by Normalization of the Acoustic Space. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 893–896, Toronto, Canada, May 1991.

[Bahl & Jelinek$^+$ 83] L. R. Bahl, F. Jelinek, R. L. Mercer: A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, pp. 179–190, March 1983.

[Baker 75] J. K. Baker: Stochastic Modeling for Automatic Speech Understanding. In D. R. Reddy (Ed.), *Speech Recognition*, Academic Press, New York, pp. 512–542, 1975.

[Bakis 76] R. Bakis: Continuous Speech Word Recognition via Centisecond Acoustic States. Proc. *ASA Meeting*, Washington, DC, April 1976.

[Balchandran & Mammone 98] R. Balchandran, R. J. Mammone, 1998: Non-Parametric Estimation and Correction on Non-Linear Distortion in Speech Systems. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 749–752, Seattle, WA, May 1998.

[Ballard & Brown 82] D. H. Ballard, C. M. Brown: *Computer Vision*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

[Bellman 57] R. E. Bellman: *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[Beulen 99] K. Beulen: Phonetische Entscheidungsbäume für die automatische Spracherkennung mit großem Vokabular. Ph.D. Thesis, RWTH Aachen, Computer Science Department, Aachen, Germany, 1999.

[Bocchieri 93] E. Bocchieri: Vector Quantization for the Efficient Computation of Continuous Density Likelihoods. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 692–695, Minneapolis, MN, April 1993.

[Brown & Della Pietra$^+$ 92] P. Brown, V. Della Pietra, P. de Souza, J. Lai, R. Mercer: Class-Based n-gram Models of Natural Language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[Burger & Weilhammer+ 00] S. Burger, K. Weilhammer, F. Schiel, H. G. Tillmann: Verb-Mobil Data Collection and Annotation. In: W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag: Berlin, Heidelberg, New York, pp. 537–549, 2000.

[Casacuberta & Llorens+ 01] F. Casacuberta, D. Llorens, C. Martinez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal, J. M. Vilar: Speech-To-Speech Translation based on Finite-State Transducers. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 613–616, Salt Lake City, UT, May 2001.

[Chengalvarayan 99] R. Chengalvarayan: Robust Energy Normalization using Speech/Nonspeech Discriminator for German Connected Digit Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. I, pp. 61–64, Budapest, Hungary, Sept. 1999.

[Chu & Jie+ 97] Y. C. Chu, C. Jie, V. Tung, B. Lin, R. Lee: Normalization of Speaker Variability by Spectrum Warping for Robust Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. III, pp. 1127–1130, Rhodes, Greece, Sept. 1997.

[Cox 00] S. Cox: Speaker Normalization in the MFCC Domain. Proc. *Int. Conf. on Spoken Language Processing*, Vol. II, pp. 853–856, Bejing, China, Oct. 2000.

[Davis & Mermelstein 80] S. B. Davis, P. Mermelstein: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. 28, No. 4, pp. 357–366, Aug. 1980.

[Dempster & Laird+ 77] A. P. Dempster, N. M. Laird, D. B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal Royal Statistical Society*, Series B, Vol. 39, No. 1, pp. 1–38, 1977.

[Dharanipragada & Padmanabhan 00] S. Dharanipragada, M. Padmanabhan: A Nonlinear Unsupervised Adaptation Technique for Speech Recognition. Proc. *Int. Conf. on Spoken Language Processing*, Vol. VI, pp. 556–559, Bejing, China, Oct. 2000.

[di Carlo 00] A. di Carlo: Speech-Input Corpus Acquisition. In: Final Report of the European Union ESPRIT LTR Project No. 30268, EuTrans, pp. 9–10, Sept. 2000.

[Dolfing 00] J. G. A. Dolfing: Exhaustive Search for Lower-Bound Errror-Rates in Vocal Tract Length Normalization. Proc. *Int. Conf. on Spoken Language Processing*, Vol. I, pp. 762–765, Bejing, China, Oct. 2000.

[Duda & Hart 73] R. O. Duda, P. E. Hart: *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[Eide & Gish 96] E. Eide, H. Gish: A Parametric Approach to Vocal Tract Length Normalization. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 346–349, Atlanta, GA, May 1996.

[Emori & Shinoda 01] T. Emori, K. Shinoda: Rapid Vocal Tract Length Normalization using Maximum Likelihood Estimation. Proc. *European Conf. on Speech Communication and Technology*, Vol. III, pp. 1649–1652, Aalborg, Denmark, Sept. 2001.

[Faltlhauser & Pfau$^+$ 00] R. Faltlhauser, T. Pfau, G. Ruske: On-line Speaking Rate Estimation Using Gaussian Mixture Models. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. III, pp. 1355–1358, Istanbul, Turkey, June 2000.

[Fritsch 97] J. Fritsch: ACID/HNN: A Framework for Hierarchical Connectionist Acoustic Modeling. Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 164–171, Santa Barbara, CA, Dec. 1997.

[Gales 01] M. J. F. Gales: Adaptive Training for Robust ASR. Proc. *IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento, Italy, 6 pages, CD ROM, IEEE Catalog No. 01EX544, Dec. 2001.

[Giuliani 99] D. Giuliani: An On-Line Acoustic Compensation Technique for Robust Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. VI, pp. 2487–2490, Budapest, Hungary, Sept. 1999.

[Gopinath 00] R. A. Gopinath: Gaussianization. *IMA Workshop: Mathematical Foundations of Speech Processing and Recognition*, Minneapolis, MN, Sept. 2000.

[Gouvêa & Stern 97] E. B. Gouvêa, R. M. Stern: Speaker Normalization through Formant-Based Warping of the Frequency Scale. Proc. *European Conf. on Speech Communication and Technology*, Vol. III, pp. 1139–1142, Rhodes, Greece, Sept. 1997.

[Haeb-Umbach 99] R. Haeb-Umbach: Investigations on Inter-Speaker Variability in the Feature Space. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 397–400, Phoenix, AZ, March 1999.

[Hermansky 90] H. Hermansky: Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal Acoustical Society of America*, Vol. 87, No. 4, pp. 1738–1752, April 1989.

[Hilger & Ney 01] F. Hilger, H. Ney: Quantile Based Histogram Equalization for Noise Robust Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. II, pp. 1135–1138, Aalborg, Denmark, Sept. 2001.

[Hilger & Molau$^+$ 02] F. Hilger, S. Molau, H. Ney: Quantile Based Histogram Equalization For Online Applications. Proc. *Int. Conf. on Spoken Language Processing*, Vol. I, pp. 237–240, Denver, CO, Sept. 2002.

[Hon & Lee 91] H. W. Hon, K. F. Lee: Recent Progress in Robust Vocabulary-Independent Speech Recognition. Proc. *DARPA Speech and Natural Language Processing Workshop*, pp. 258-263, Pacific Grove, Feb. 1991.

[Huang & Jack 89] X. D. Huang, M. A. Jack: Semi-Continuous Hidden Markov Models for Speech Signals. *Computer Speech and Language*, Vol. 3, No. 3, pp. 329–252, 1989.

[Huang & Acero+ 95] X. Huang, A. Acero, F. Alleva, M. Y. Hwang, L. Jiang, M. Mahajan: Microsoft Windows Highly Intelligent Speech Recognizer: WHISPER. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 93–96, Adelaide, Australia, April 1995.

[Hung & Wang 00] W. W. Hung, H. C. Wang: A Fuzzy Approach for the Equalization of Cepstral Variances. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. III, pp. 1611–1614, Istanbul, Turkey, June 2000.

[Hwang & Huang+ 92] M. Y. Hwang, X. D. Huang, F. Alleva: Predicting Unseen Triphones with Senones. Technical Report No. 510.7808 C28R 93-139 2, Carnegie Mellon University, Pittsburgh, PA, 1993.

[Jelinek 69] F. Jelinek: A Fast Sequential Decoding Algorithm Using a Stack. *IBM Journal of Research and Development*, Vol. 13, pp. 675–685, Nov. 1969.

[Jelinek 76] F. Jelinek: Continuous Speech Recognition by Statistical Methods. *Proc. of the IEEE*, Vol. 64, No. 10, pp. 532–556, April 1976.

[Jelinek 91] F. Jelinek: Self-Organized Language Modeling for Speech Recognition. In: A. Waibel, K.-F. Lee (Eds.), *Readings in Speech Recognition*, Morgan Kaufmann Publ., San Mateo, CA, pp. 450–506, 1991.

[Junqua & Fohr+ 95] J.-C. Junqua, D. Fohr, J.-F. Mari, T. H. Applebaum, B. A. Hanson: Time Derivatives, Cepstral Normalization, and Spectral Parameter Filtering for Continuously Spelled Names over Telephone. Proc. *European Conf. on Speech Communication and Technology*, Vol. II, pp. 1385–1388, Madrid, Spain, Sept. 1995.

[Junqua & Haton 99] J.-C. Junqua, J. P. Haton: *Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[Kamm & Andreou+ 95] T. Kamm, G. Andreou, J. Cohen: Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability. Proc. *15th Annual Speech Research Symposion*, pp. 175–178, CSLP, Johns Hopkins University, Baltimore, MD, June 1995.

[Kanthak & Schütz+ 00] S. Kanthak, K. Schütz, H. Ney: Using SIMD Instructions For Fast Likelihood Calculation in LVCSR. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. III, pp. 1531–1534, Istanbul, Turkey, June 2000.

[Kanthak & Sixtus+ 00] S. Kanthak, A. Sixtus, S. Molau, R. Schlüter, H. Ney: Fast Search for Large Vocabulary Speech Recognition. In: W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag: Berlin, Heidelberg, New York, pp. 63–78, 2000.

[Katz 87] S. M. Katz: Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 35, No. 3, pp. 400–401, March 1987.

[Kubala 95] F. Kubala: Design of the 1994 CSR Benchmark Tests. Proc. *ARPA Spoken Language Technology Workshop*, pp. 41–46, Austin, TX, Jan. 1995.

[Kuhn & de Mori 90] R. Kuhn, R. de Mori: A Cache-Based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 6, pp. 570–583, June 1990.

[Lee & Rose 96] L. Lee, R. Rose: Speaker Normalization using Efficient Frequency Warping Procedures. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 353–356, Atlanta, GA, May 1996.

[Leggetter & Woodland 95] C. J. Leggetter, P. C. Woodland: Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer, Speech and Language*, Vol. 9, pp. 171–185, April 1995.

[Levinson & Rabiner$^+$ 83] S. E. Levinson, L. R. Rabiner, M. M. Sondhi: An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Techn. Journal*, Vol. 62, No. 4, pp. 1035–1074, April 1983.

[Liu & Acero$^+$ 92] F. H. Liu, A. Acero, R. M. Stern: Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 257–260, San Francisco, CA, March 1992.

[Macherey & Ney 02] W. Macherey, H. Ney: Towards Automatic Corpus Preparation for a German Broadcast News Transcription System. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. I, pp. 733–736, Orlando, Florida, May 2002.

[Martin & Hamacher$^+$ 99] S. Martin, C. Hamacher, J. Liermann, F. Wessel, H. Ney: Assessment of Smoothing Methods and Complex Stochastic Language Modeling. Proc. *European Conf. on Speech Communication and Technology*, Vol. V, pp. 1939–1942, Budapest, Hungary, Sept. 1999.

[Mashao 96] D. J. Mashao: Experiments on a Parametric Nonlinear Spectral Warping for an HMM-Based Speech Recognizer. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 17–21, Atlanta, GA, May 1996.

[Matsukoto & Hirowo 92] H. Matsukoto, I. Hirowo: A Piecewise Linear Mapping for Supervised Speaker Adaptation. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 449–452, San Francisco, CA, March 1992.

[McDonough & Byrne$^+$ 98] J. McDonough, W. Byrne, X. Luo: Speaker Normalization with All-Pass Transforms. Proc. *Int. Conf. on Spoken Language Processing*, Vol. VI, pp. 2307–2310, Sydney, Australia, Nov. 1998.

[Mirghafori & Fosler$^+$ 95] N. Mirghafori, E. Fosler, N. Morgan: Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes. Proc. *European Conf. on Speech Communication and Technology*, Vol. I, pp. 491–494, Madrid, Spain, Sept. 1995.

[Mokbel & Monnè+ 93] C. Mokbel, J. Monnè, D. Jouvet: On-line Adaptation of a Speech Recognizer to Variations in Telephone Line Conditions. Proc. *European Conf. on Speech Communication and Technology*, Vol. II, pp. 1247–1250, Berlin, Germany, Sept. 1993.

[Molau & Kanthak+ 00] S. Molau, S. Kanthak, H. Ney: Efficient Vocal Tract Normalization in Automatic Speech Recognition. Proc. *Elektronische Sprachsignalverarbeitung*, pp. 209–216, Cottbus, Germany, Sept. 2000.

[Molau & Pitz+ 01a] S. Molau, M. Pitz, R. Schlüter, H. Ney: Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 73–76, Salt Lake City, UT, May 2001.

[Molau & Pitz+ 01b] S. Molau, M. Pitz, H. Ney: Histogram Based Normalization in the Acoustic Feature Space. Proc. *IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento, Italy, 4 pages, CD ROM, IEEE Catalog No. 01EX544, Dec. 2001.

[Molau & Hilger+ 02] S. Molau, F. Hilger, D. Keysers, H. Ney: Enhanced Histogram Normalization in the Acoustic Feature Space. Proc. *Int. Conf. on Spoken Language Processing*, Vol. I, pp. 1421-1424, Denver, CO, Sept. 2002.

[Naik 95] D. Naik: Pole-Filtered Cepstral Mean Subtraction. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 157–161, Detroit, MI, May 1995.

[Naito & Deng+ 99] M. Naito, L. Deng, Y. Sagisaka: Model-Based Speaker Normalization Methods for Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. VI, pp. 2515–2518, Budapest, Hungary, Sept. 1999.

[Nene & Nayar 1996] S. A. Nene, S. K. Nayar: Closest Point Search in High Dimensions. Proc. *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 859-865, San Francisco, CA, June 1996.

[Neumeyer & Weintraub 94] L. Neumeyer, M. Weintraub: Probabilistic Optimum Filtering for Robust Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 417–420, Adelaide, Australia, April 1995.

[Neumeyer & Sankar+ 95] L. Neumeyer, A. Sankar, V. Digalakis: A Comparative Study of Speaker Adaptation Techniques. Proc. *European Conf. on Speech Communication and Technology*, Vol. II, pp. 1127–1130, Madrid, Spain, Sept. 1995.

[Ney 84] H. Ney: The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 32, No. 2, pp. 263–271, April 1984.

[Ney & Mergel+ 87] H. Ney, D. Mergel, A. Noll, A. Paeseler: A Data-Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 833–836, Dallas, TX, April 1987.

[Ney & Noll 88] H. Ney, A. Noll: Phoneme Modeling using Continuous Mixture Densities. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 437–440, New York, April 1988.

[Ney 1990] H. Ney: Acoustic Modeling of Phoneme Units for Continuous Speech Recognition. In: L. Torres, E. Masgrau, M. A. Lagunas (Eds.): *Signal Processing V: Theories and Applications*, pp. 65-72, Elsevier Science Publishers, The Netherlands, 1990.

[Ney & Haeb-Umbach$^+$ 92] H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder: Improvements in Beam Search for 10000-Word Continuous Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 9–12, San Francisco, CA, March 1992.

[Ney & Essen$^+$ 94] H. Ney, U. Essen, R. Kneser: On Structuring Probabilistic Dependencies in Language Modeling. *Computer Speech and Language*, Vol. 2, No. 8, pp. 1–38, 1994.

[Ney & Welling$^+$ 98] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel: The RWTH Large Vocabulary Continuous Speech Recognition System. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 853–856, Seattle, WA, May 1998.

[Odell & Valtchev$^+$ 94] J. J. Odell, V. Valtchev, P. C. Woodland, S. J. Young: A One-Pass Decoder Design for Large Vocabulary Recognition. Proc. *ARPA Spoken Language Technology Workshop*, pp. 405–410, Plainsboro, NJ, March 1994.

[Ortmanns & Ney 1995] S. Ortmanns, H. Ney: An Experimental Study of the Search Space for 20000-Word Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. II, pp. 901–904, Madrid, Spain, Sept. 1995.

[Ortmanns & Ney$^+$ 97] S. Ortmanns, H. Ney, T. Firzlaff: Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. I, pp. 139–142, Rhodes, Greece, Sept. 1997.

[Ortmanns 1998] S. Ortmanns: *Effiziente Suchverfahren zur Erkennung kontinuierlich gesprochener Sprache*. Ph.D. Thesis, RWTH Aachen, Computer Science Department, Becker-Kuns Verlag, Aachen, Germany, 1998.

[Padmanabhan & Dharanipragada 01] M. Padmanabhan, S. Dharanipragada: Maximum Likelihood Non-linear Transformation for Environment Adaptation in Speech Recognition Systems. Proc. *European Conf. on Speech Communication and Technology*, Vol. IV, pp. 2359–2362, Aalborg, Denmark, Sept. 2001.

[Pallett & Fiscus$^+$ 93] D. Pallett, J. Fiscus, W. Fisher, J. S. Garofolo: Benchmark Tests for the DARPA Spoken Language Program. Proc. *ARPA Human Language Technology Workshop*, pp. 7–18, Princeton, NJ, March 1993.

[Pallett & Fiscus$^+$ 94] D. Pallett, J. Fiscus, W. Fisher, J. S. Garofolo, B. Lund, M. Prysbocki: 1993 Benchmark Tests for the ARPA Spoken Language Program. Proc. *ARPA Spoken Language Technology Workshop*, pp. 49–74, Princeton, NJ, March 1994.

[Pallett & Fiscus+ 95] D. Pallett, J. Fiscus, W. Fisher, J. S. Garofolo, B. Lund, M. Prysbocki: 1994 Benchmark Tests for the ARPA Spoken Language Program. Proc. *ARPA Spoken Language Technology Workshop*, pp. 5–36, Austin, TX, Jan. 1995.

[Pfau & Faltlhauser+ 99] T. Pfau, R. Faltlhauser, G. Ruske: A Combination of Speaker Normalization and Speech Rate Normalization for Automatic Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. I, pp. 299–302, Budapest, Hungary, Sept. 1999.

[Pfau & Faltlhauser+ 00] T. Pfau, R. Faltlhauser, G. Ruske: Speaker Normalization and Pronunciation Variant Modeling: Helpful Methods for Improving Recognition of Fast Speech. Proc. *Int. Conf. on Spoken Language Processing*, Vol. IV, pp. 362–365, Bejing, China, Oct. 2000.

[Pitz & Molau+ 01] M. Pitz, S. Molau, R. Schlüter, H. Ney: Vocal Tract Normalization equals Linear Transformation in Cepstral Space. Proc. *European Conf. on Speech Communication and Technology*, Vol. VI, pp. 2653–2656, Aalborg, Denmark, Sept. 2001.

[Pye & Woodland 97] D. Pye, P. C. Woodland: Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 1047–1050, Munich, Germany, April 1997.

[Rabiner & Schafer 78] L. R. Rabiner, R. W. Schafer: *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[Rabiner & Juang 86] L. Rabiner, B.-H. Juang: An Introduction to Hidden Markov Models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 3, No. 1, pp. 4–16, Jan. 1986.

[Rabiner 89] L. R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, Vol. 77, No. 2, pp. 257–286, Feb. 1989.

[Richardson & Hwang+ 99] M. Richardson, M. Hwang, A. Acero, X. D. Huang: Improvements on Speech Recognition for Fast Talkers. Proc. *European Conf. on Speech Communication and Technology*, Vol. I, pp. 411–414, Budapest, Hungary, Sept. 1999.

[Rosenfeld 94] R. Rosenfeld: *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*, Ph.D. Thesis, Technical Report CMU-CS-94-138, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1994.

[Sankar & Lee 95] A. Sankar, C.-H. Lee: Robust Speech Recognition Based on Stochastic Matching. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 121–124, Adelaide, Australia, April 1995.

[Schwartz & Chow+ 85] R. Schwartz, Y.-L. Chow. O. Kimball, S. Roucos, U. Krasner, J. Makhoul: Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 3, pp. 1205–1208, Tampa, FL, March 1985.

[Schwartz & Chow 90] R. Schwartz, Y.-L. Chow: The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N most likely Sentence Hypotheses. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 81–84, Albuquerque, NM, April 1990.

[Schwartz & Austin 91] R. Schwartz, S. Austin: A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 701–704, Toronto, May 1991.

[Siegler & Stern 95] M. A. Siegler, R. M. Stern: On the Effect of Speech Rate in Large Vocabulary Speech Recognition Systems. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 612–615, Adelaide, Australia, April 1995.

[Sixtus & Molau$^+$ 00] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, H. Ney: Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. III, pp. 1671–1674, Istanbul, Turkey, June 2000.

[Sixtus 02] A. Sixtus: Across-Word Phoneme Models for Large Vocabulary Continuos Speech Recognition. Ph.D. Thesis (in preparation), RWTH Aachen, Computer Science Department, Aachen, Germany, 2002.

[Steinbiss & Tran$^+$ 94] V. Steinbiss, B.-H. Tran, H. Ney: Improvements in Beam Search. Proc. *Int. Conf. on Spoken Language Processing*, Vol. IV, pp. 2143–2146, Yokohama, Sept. 1994.

[Uebel & Woodland 99] L. F. Uebel, P. C. Woodland: An Investigation into Vocal Tract Length Normalisation. Proc. *European Conf. on Speech Communication and Technology*, Vol. VI, pp. 2527–2530, Budapest, Hungary, Sept. 1999.

[van Kampen 92] N. G. van Kampen: *Stochastic Processes in Physics and Chemistry*. Elsevier Science Publishers, Amsterdam, The Netherlands, 1992.

[van Compernolle 89] D. van Compernolle: Noise Adaptation in a Hidden Markov Model Speech Recognition System. *Computer, Speech and Language*, Vol. 3, No. 2, pp. 151–167.

[Vintsyuk 71] T. K. Vintsyuk: Elementwise Recognition of Continuous Speech Composed of Words from a Specified Dictionary. *Cybernetics*, Vol. 7, pp. 133–143, March 1971.

[Viterbi 67] A. Viterbi: Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Transactions on Information Theory*, Vol. 13, No. 2, pp. 260–269, 1967.

[Wakita 77] H. Wakita: Normalization of Vowels by Vocal-Tract Length and its Application to Vowel Identification. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 25, pp. 183–192, April 1977.

[Wahlster 00] W. Wahlster (Ed.): *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer Verlag: Berlin, Heidelberg, New York, 2000.

[Wegmann & McAllaster+ 96] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin: Speaker Normalization on Conversational Telephone Speech. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 339–342, Atlanta, GA, May 1996.

[Welling & Haberland+ 97] L. Welling, N. Haberland, H. Ney: Acoustic Front-End Optimization for Large Vocabulary Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. IV, pp. 2099–2102, Rhodes, Greece, Sept. 1997.

[Welling & Haeb-Umbach+ 98] L. Welling, R. Haeb-Umbach, X. Aubert, N. Haberland: A Study on Speaker Normalisation using Vocal Tract Normalisation and Speaker Adaptive Training. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 797–800, Seattle, WA, May 1998.

[Welling & Kanthak+ 99] L. Welling, S. Kanthak, H. Ney: Improved Methods for Vocal Tract Normalization. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 761–764, Phoenix, AZ, March 1999.

[Welling 99] L. Welling: Merkmalsextraktion in Spracherkennungssystemen für großen Wortschatz. Ph.D. Thesis, RWTH Aachen, Computer Science Department, Aachen, Germany, 1999.

[Wessel & Ortmanns+ 97] F. Wessel, S. Ortmanns, H. Ney: Implementation of Word Based Statistical Language Models. Proc. *SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, pp. 55-59, Pilsen, Czech Republic, April 1997.

[Wessel & Ney 01] F. Wessel, H. Ney: Unsupervised Training for Broadcast News Speech Recognition. Proc. *IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento, Italy, 4 pages, CD ROM, IEEE Catalog No. 01EX544, Dec. 2001.

[Westphal 97] M. Westphal: The Use of Cepstral Means in Conversational Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. III, pp. 1143–1146, Rhodes, Greece, Sept. 1997.

[Westphal & Schultz+ 98] M. Westphal, T. Schultz, A. Waibel: Linear Discriminant - a new Criterion for Speaker Normalization. Proc. *Int. Conf. on Spoken Language Processing*, Vol. III, pp. 827–830, Sydney, Australia, Nov. 1998.

[Young 93] S. J. Young: HTK: Hidden Markov model toolkit V1.4. User Manual, Cambridge University Engineering Department, Cambridge, England, Feb. 1993.

[Yu & Waibel 00] H. Yu, A. Waibel: Streamlining the Front-End of a Speech Recognizer. Proc. *Int. Conf. on Spoken Language Processing*, Vol. I, pp. 353–356, Bejing, China, Oct. 2000.

[Zhan & Westphal 97] P. Zhan, M. Westphal: Speaker Normalization based on Frequency Warping. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 1039–1042, Munich, Germany, April 1997.

[Zhu & Alwan 00] Q. Zhu, A. Alwan: On the Use of Variable Frame Rate Analysis in Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. III, pp. 1783–1786, Istanbul, Turkey, June 2000.

Appendix A

Symbols and Acronyms

A.1 Mathematical Symbols

$\mathbf{X}$	sequence of (unnormalized) acoustic vectors $x_1, \ldots, x_T$
$\tilde{\mathbf{X}}$	sequence of normalized acoustic vectors $\tilde{x}_1, \ldots, \tilde{x}_T$
$\mathbf{X}^\alpha$	sequence of acoustic vectors obtained by warping the frequency axis with the warping factor α
$\mathbf{X_r}$	sequence of acoustic vectors from condition r
$\mathbf{W}$	sequence w_1^N of spoken words $w_1, \ldots, w_N$
$\hat{\mathbf{W}}$	sequence of recognized words from a first recognition pass
$\mathbf{S}$	sequence of hidden Markov model states
$\mathbf{T}$	number of time frames
$\mathbf{L}$	number of densities in a Gaussian mixture
$\mathbf{N}$	number of spoken words
$\mathbf{V}$	vocabulary size
$\mathbf{K}$	number of cepstrum coefficients
$\mathbf{D}$	dimension of the acoustic vector
$\mathbf{R}$	number of acoustic conditions
$\mathbf{f}_\alpha$	generic transformation function with parameter set α
θ	set of all parameters of an (unnormalized) acoustic model
$\tilde{\theta}$	set of all parameters of a normalized acoustic model
$\mathbf{p}(\mathbf{W})$	language model or a-priori probability of the word sequence W

$\mathbf{p}(\mathbf{X})$	a-priori probability of the sequence of acoustic vectors X
$\mathbf{p}(\mathbf{X}\|\mathbf{W})$	acoustic probability of the acoustic vector sequence X given the word sequence W
$\tilde{\mathbf{p}}(\mathbf{X}\|\mathbf{W})$	adapted acoustic probability
$\mathbf{p}(\mathbf{x}\|\mathbf{s},\mathbf{W};\theta)$	hidden Markov model emission probability for the acoustic vector x at state s
$\mathbf{p}(\mathbf{s_t}\|\mathbf{s_{t-1}},\mathbf{W})$	hidden Markov model transition probability for the transition from state s_{t-1} to state s_t
$\mathbf{s_1^T}$	set of all possible alignments between acoustic vectors X and hidden Markov model states S
$\mathcal{N}(\mu,\mathbf{\Sigma})$	normal (Gaussian) distribution with mean vector μ and covariance matrix Σ
$\mathbf{c_l}$	mixture weight for the density with index l
$\theta_\mathbf{0}$	unnormalized low resolution acoustic model
$\tilde{\theta}_\mathbf{0}$	normalized low resolution acoustic model
$\mathbf{\Lambda}_\alpha$	Gaussian mixture model trained on unnormalized data from speakers with warping factor α
$\tilde{\mathbf{\Lambda}}$	Gaussian mixture model trained on normalized data
$\mathbf{p}(\mathbf{X}\|\mathbf{\Lambda})$	acoustic probability of the acoustic vector sequence X given the Gaussian mixture model Λ
$\alpha_\mathbf{r}$	frequency axis warping factor of speaker r
$\hat{\alpha}_\mathbf{r}$	estimated frequency axis warping factor
ω	frequency
$\tilde{\omega}$	normalized (warped) frequency
$\omega_\mathbf{0}$	limiting frequency for piece-wise linear frequency axis warping
$\mathbf{z}(\mathbf{x})$	weight for the acoustic vector x
$\mathbf{e}(\mathbf{x})$	energy of the acoustic vector x
$\mathbf{m}$	log-compressed magnitude spectrum
$\mathbf{c}$	vector of cepstral coefficients $c_0,\ldots,c_{K-1}$
$\mathbf{U}$	cepstrum transformation matrix

$\mathbf{g}(\omega)$	arbitrary monotone invertible frequency axis warping function
$\mu(\omega)$	Mel-frequency warping function
$\tilde{\mu}(\omega)$	normalized Mel-frequency warping function
$\nu_\alpha(\omega)$	VTN frequency warping function, depending on the warping factor α
$\chi(\omega)$	combined VTN and Mel-frequency warping function
$\mathbf{f_s}$	sampling frequency
β	alternative warping factor for piece-wise linear warping
κ	alternative warping factor for piece-wise linear warping
$\mathbf{p_r}(\mathbf{x})$	probability distribution or histogram of condition r
$\mathbf{P_r}(\mathbf{x})$	cumulative probability distribution or histogram of condition r
$\tilde{\mathbf{p}}(\mathbf{x})$	reference probability distribution or histogram
$\tilde{\mathbf{P}}(\mathbf{x})$	cumulative reference probability distribution or histogram
$\tilde{\mathbf{p}}_{\mathbf{sil}}(\mathbf{x})$	reference histogram estimated on silence frames only
$\tilde{\mathbf{P}}_{\mathbf{sil}}(\mathbf{x})$	cumulative reference histogram estimated on silence frames only
$\tilde{\mathbf{p}}_{\mathbf{sp}}(\mathbf{x})$	reference histogram estimated on speech frames only
$\tilde{\mathbf{P}}_{\mathbf{sp}}(\mathbf{x})$	cumulative reference histogram estimated on speech frames only
$\tilde{\mathbf{p}}_{\mathbf{r}}(\mathbf{x})$	reference histogram adapted to the silence fraction of condition r
$\tilde{\mathbf{P}}_{\mathbf{r}}(\mathbf{x})$	cumulative reference histogram adapted to the silence fraction of condition r
$\gamma_{\mathbf{r}}$	silence fraction of condition r
$\mathbf{\Sigma}$	condition-dependent covariance matrix
$\tilde{\mathbf{\Sigma}}$	reference covariance matrix
$\mathbf{v_d}$	condition-dependent eigenvector for dimension d
$\hat{\mathbf{v}}_{\mathbf{d}}$	transformed condition-dependent eigenvector for dimension d
$\mathbf{V}$	matrix made of the condition-dependent eigenvectors
$\tilde{\mathbf{v}}_{\mathbf{d}}$	reference eigenvector for dimension d
$\tilde{\mathbf{V}}$	matrix made of the reference eigenvectors
$\lambda_{\mathbf{d}}$	condition-dependent eigenvalue for dimension d

$\tilde{\lambda}_{\mathbf{d}}$	reference eigenvalue for dimension d
$\eta_{\mathbf{d}}$	rotation angle to match the eigenvectors for dimension d
$\mathbf{I}$	identity matrix
$\mathbf{U}_{\mathbf{d}}$	transformation matrix for acoustic vectors to match d eigenvectors
$\hat{\mathbf{U}}_{\mathbf{d}}$	transformation matrix for dimension d
$\mathbf{J}_{\mathbf{d}}$	projection matrix for dimension d
$\mathbf{R}_{\mathbf{d}}$	rotation matrix for dimension d

A.2 Acronyms

RWTH	**R**heinisch-**W**estfälische **T**echnische **H**ochschule
LVCSR	**l**arge **v**ocabulary **c**onversational **s**peech **r**ecognition
OOV	**o**ut **o**f **v**ocabulary
ML	**m**aximum **l**ikelihood
EM	**e**xpectation **m**aximization
HMM	**h**idden **M**arkov **m**odel
GMM	**G**aussian **m**ixture **m**odel
LDA	**l**inear **d**iscriminant **a**nalysis
MFCC	**M**el-**f**requency **c**epstral **c**oefficients
PLP	**p**erceptual **l**inear **p**rediction
RASTA	**r**el**a**tive **s**pec**t**r**a**l
CMN	**c**epstral **m**ean **n**ormalization
CVN	**c**epstral **v**ariance **n**ormalization
VTN	**v**ocal tract **l**ength **n**ormalization
MLLR	**m**aximum **l**ikelihood **l**inear **r**egression
MAP	**m**aximum **a**-**p**osteriori
ROS	**r**ate **o**f **s**peech
LPC	**l**inear **p**redictive **c**oding
CART	**c**lassification **a**nd **r**egression **t**ree
FFT	**f**ast **F**ourier **t**ransform
DCT	**d**iscrete **c**osine **t**ransform
SNR	**s**ignal to **n**oise **r**atio
PCA	**p**rincipal **c**omponent **a**nalysis
SIMD	**s**ingle **i**nstruction, **m**ultiple **d**ata
LM	**l**anguage **m**odel
PP	**p**er**p**lexity
WER	**w**ord **e**rror **r**ate

DEL	**del**etion errors
INS	**ins**ertion errors
DEV	**dev**elopment test corpus
EVAL	**eval**uation test corpus
RTF	**r**eal-**t**ime **f**actor
WSJ	**W**all **S**treet **J**ournal - a speech corpus
NAB	**N**orth **A**merican **B**usiness **N**ews - a speech corpus
TIDIGITS	**T**exas **I**nstruments connected **digit s**equences - a speech corpus
TIMIT	speech corpus collected by **T**exas **I**nstruments and transcribed at the **M**assachusetts **I**nstitute of **T**echnology
SIETILL	speech corpus collected by **Sie**mens and the Institute of Phonetics of Prof. **Till**mann at the University of Munich

Lebenslauf - Curriculum Vitae

Name: Molau

Vorname: Sirko

Geburtsdatum: 04.03.1971

Geburtsort: Berlin

Eltern: Hans-Joachim Molau, Diplom-Ingenieur,
Eva Molau geb. Kernbach, Diplom-Ingenieurökonom

Staatsangehörigkeit: Bundesrepublik Deutschland

Schulausbildung:

- 10 Jahre Polytechnische Oberschule
- 2 Jahre Erweiterte Oberschule

Studium:

- 1990-93 an der Technischen Universität Chemnitz, Sektion Informatik
- 1993-94 an der University of York / England, Department of Computer Science
- 1994-97 an der Technischen Universität Chemnitz, Fakultät für Informatik
- Juli 1995 an der University of Illinois at Urbana-Champaign / USA, Intensive English Institute
- 1997-2002 an der RWTH Aachen, Lehrstuhl für Informatik VI

Abgelegte Prüfungen:

- 1987 Mittlere Reifeprüfung, Abschluß "mit Auszeichnung"
- 1989 Abitur, Abschluß "mit Auszeichnung"
- 1992 Vordiplom Informatik
- 1997 Diplom Informatik, Abschluß mit "sehr gut"
- 2003 Promotion Informatik, Abschluß mit "sehr gut"

Berufstätigkeit:

- 1987-1997 nebenberuflich freier Mitarbeiter an der Archenhold-Sternwarte Berlin
- 1989-1990 Wehrdienst
- 1990-1997 Student
- 1997 wiss. Hilfskraft am Institut für Planetenerkundung der Deutschen Forschungsanstalt für Luft- und Raumfahrt
- 1997-2002 wiss. Mitarbeiter am Lehrstuhl für Informatik VI der RWTH Aachen
- seit 2002 IT-Spezialist bei der BMW AG