Investigations on Linear Transformations for Speaker Adaptation and Normalization

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

von

Diplom–Physiker Michael Pitz

aus

Aachen

Berichter: Universitätsprofessor Dr.-Ing. Hermann Ney
Professor Dr. Christian Wellekens

Tag der mündlichen Prüfung: 14. März 2005

Zwei Dinge sind zu unserer Arbeit nötig: Unermüdliche Ausdauer und die Bereitschaft, etwas, in das man viel Zeit und Arbeit gesteckt hat, wieder wegzuwerfen.

Albert Einstein

Acknowledgments

First I would like to thank my supervisor Prof. Dr.-Ing. Hermann Ney, head of the Lehrstuhl für Informatik VI at the RWTH Aachen, for the opportunity to realize this work as part of your team. You introduced me to the exciting field of pattern recognition in general and speech recognition in particular. You allowed me great latitude to pursue my ideas and followed them with great interest. I would also like to thank you for the numerous interesting and enlightening discussions we had.

I am also grateful to my second supervisor Prof. Christian Wellekens, who is with the Multimedia Communications Department of Institut Eurécom, France, for your interest in my work, the in-depth reading of this thesis and the valuable comments.

Stephan Kanthak, you have been an enormous help in many computer problems and difficult debugging sessions. I always admired your deep insight in computer technology, Linux and C++. Besides that, we had many funny talks about the world and his brother.

Ralf Schlüter, I am grateful for our discussions and numerous sessions at the whiteboard, which gave me a deeper insight into speech recognition and helped to solve a couple of problems.

Oliver Bender, Michael Motter, Stefan Koltermann, Mirko Kohns, Achim Sixtus and Klaus Macherey, you kept the computers running and patiently dealt with all my requests.

I always enjoyed very much the relaxing time at lunch and coffee breaks with the "Geigeltruppe" Achim, Andras, Frank, Nicola, Ralf, Sirko, Sonja, and Stephan.

To all current and former colleagues of the Lehrstuhl für Informatik VI for the motivating atmosphere, many interesting discussions and also many laughter.

I want to express a very special thank to my girlfriend Beate. You had an important part in the success of this thesis. Without you, life would be less wonderful.

Nicht zuletzt möchte ich besonders meinen Eltern danken. Ihr habt meinen Weg immer verfolgt, mich ermutigt und unterstützt.

This work was partially funded by the European Commission under the Human Language Technologies project CORETEX (IST-1999-11876), and by the DFG (Deutsche Forschungsgemeinschaft) under contract NE 572/4-1 and NE 572/4-3.

Abstract

This thesis deals with linear transformations at various stages of the automatic speech recognition process.

In current state-of-the-art speech recognition systems linear transformations are widely used to care for a potential mismatch of the training and testing data and thus enhance the recognition performance. A large number of approaches has been proposed in literature, though the connections between them have been disregarded so far. By developing a unified mathematical framework, close relationships between the particular approaches are identified and analyzed in detail.

Mel frequency Cepstral coefficients (MFCC) are commonly used features for automatic speech recognition systems. The traditional way of computing MFCCs suffers from a twofold smoothing, which complicates both the MFCC computation and the system optimization. An improved approach is developed that does not use any filter bank and thus avoids the twofold smoothing. This integrated approach allows a very compact implementation and needs less parameters to be optimized.

Starting from this new computation scheme for MFCCs, it is proven analytically that *vocal tract normalization (VTN)* equals a linear transformation in the Cepstral space for arbitrary invertible warping functions. The transformation matrix for VTN is explicitly calculated exemplary for three commonly used warping functions. Based on some general characteristics of typical VTN warping functions, a common structure of the transformation matrix is derived that is almost independent of the specific functional form of the warping function. By expressing VTN as a linear transformation it is possible, for the first time, to take the Jacobian determinant of the transformation into account for any warping function. The effect of considering the Jacobian determinant on the warping factor estimation is studied in detail.

The second part of this thesis deals with a special linear transformation for speaker adaptation, the *Maximum Likelihood Linear Regression (MLLR)* approach. Based on the close interrelationship between MLLR and VTN proven in the first part, the general structure of the VTN matrix is adopted to restrict the MLLR matrix to a band structure, which significantly improves the MLLR adaptation for the case of limited available adaptation data.

Finally, several enhancements to MLLR speaker adaptation are discussed. One deals with refined definitions of regression classes, which is of special importance for fast adaptation when only limited adaptation data are available. Another enhancement makes use of confidence measures to care for recognition errors that decrease the adaptation performance in the first pass of a two-pass adaptation process.

Zusammenfassung

Diese Arbeit befaßt sich mit linearen Transformationen an verschiedenen Stellen des automatischen Spracherkennungsprozesses.

In modernen automatischen Spracherkennungssystemen sind lineare Transformationen ein beliebtes Mittel, um einer Diskrepanz von Trainings- und Testdaten entgegenzuwirken und somit die Erkennungsleistung zu steigern. Eine Vielzahl von Ansätzen ist in der Literatur vorgeschlagen worden, allerdings wurden die Zusammenhänge zwischen den Ansätzen bisher vernachlässigt. Durch die Entwicklung einer vereinheitlichten mathematischen Beschreibung werden enge Zusammenhänge zwischen den einzelnen Ansätzen aufgezeigt und ausführlich untersucht.

Mel-Frequenz Cepstrum Koeffizienten (MFCC) werden sehr häufig als Merkmale in automatischen Spracherkennungssystemen eingesetzt. Der übliche Ansatz zur Berechnung der MFCC beinhaltet allerdings eine doppelte Glättung, was sowohl die Berechnung der MFCC als auch die Parameteroptimierung erschwert. Es wird ein verbesserter Ansatz vorgestellt, der auf eine Filterbank verzichtet und somit die doppelte Glättung vermeidet. Dieser integrierte Ansatz erlaubt eine sehr kompakte Implementierung und benötigt weniger zu optimierende Parameter.

Ausgehend von dieser neuen Methode zur Berechnung der MFCC wird analytisch gezeigt, daß *Vokaltraktlängennormierung (VTN)* für beliebige invertierbare Verzerrungsfunktionen als eine lineare Transformation im Cepstrumraum dargestellt werden kann. Die Transformationsmatrix für VTN wird beispielhaft für drei häufig verwendete Verzerrungsfunktionen explizit berechnet. Basierend auf einigen generellen Eigenschaften typischer VTN Verzerrungsfunktionen wird eine gemeinsame Struktur der Transformationsmatrizen abgeleitet, die größtenteils unabhängig von der funktionellen Form der Verzerrungsfunktion ist. Durch die Möglichkeit VTN als lineare Transformation auszudrücken ist es erstmals möglich die Jacobi-Determinante der Transformation für beliebige Warpingfunktionen zu berücksichtigen. Die Auswirkungen der Berücksichtigung der Jacobi-Determinante bei der Warpingfaktorschätzung werden ausführlich untersucht.

Der zweite Teil dieser Arbeit beschäftigt sich mit einer speziellen linearen Transformation zur Sprecheradaption, des *Maximum Likelihood Linear Regression (MLLR)* Ansatzes. Basierend auf dem engen Zusammenhang von MLLR und VTN, der im ersten Teil gezeigt wurde, wird die generelle Form der VTN-Matrix auf die MLLR-Matrix übertragen, um diese auf eine Bandstruktur einzuschränken. Dadurch wird die MLLR Adaption besonders für den Fall von wenigen verfügbaren Adaptionsdaten erheblich verbessert.

Schließlich werden mehrere Verbesserungen der Sprecheradaption mittels MLLR präsentiert. Eine Erweiterung zielt auf eine verbesserte Definition der Regressionsklassen ab, was speziell für den Fall einer schnellen Adaption mit wenigen Adaptionsdaten eine besondere Bedeutung hat. Eine weitere Verbesserung nutzt Konfidenzmaße, um einer Verschlechterung der Adaptionsleistung durch Erkennungsfehler im ersten Durchgang eines mehrstufigen Adaptionsprozesses entgegenzuwirken.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistical Speech Recognition

In recent years the statistical approach to speech recognition has prevailed over other approaches. Given a sequence of acoustic observations $x_1^T = x_1, \ldots, x_T$, that word sequence $w_1^N = w_1, \ldots, w_N$ should be chosen according to Bayes' decision rule which maximizes the *a-posteriori* probability [Bayes 1763]:

$$\begin{aligned} [w_1^N]_{\text{opt}} &= \underset{w_1^N}{\text{argmax}}\, p(w_1^N|x_1^T) \\ &= \underset{w_1^N}{\text{argmax}} \left\{ p(x_1^T|w_1^N) \cdot p(w_1^N) \right\} \end{aligned} \tag{1.1}$$

Eq. (1.1) shows the two basic stochastic models that are involved in automatic speech recognition: the *acoustic model* $p(x_1^T|w_1^N)$, i.e. the probability of observing the sequence of feature vectors x_1^T given a word sequence w_1^N, and the *language model* $p(w_1^N)$ which provides an a-priori probability for a word sequence w_1^N. The basic architecture of a statistical speech recognition system is depicted in Figure 1.1 [Ney 90]. The system consists of four main components which will be described in detail in the following Sections:

- the *signal analysis* (Section 1.2) module aims at extracting acoustic features from the input speech signal and provides the speech recognizer with a sequence of acoustic vectors x_1^T

- the *acoustic model* (Section 1.3) consists of statistical models for the smallest sub-words units to be distinguished by the speech recognizer, e.g. phonemes, syllables or whole words, and a pronunciation lexicon which defines the composition of an acoustic model for a given word from the sub-word units

- the *language model* (Section 1.4) provides the a-priori probability of a hypothesized word sequence based on the syntax, semantics and pragmatics of the language to be recognized

- the *search module* (Section 1.5) finally combines the two knowledge sources acoustic model and language model and determines the word sequence that maximizes Eq. (1.1).

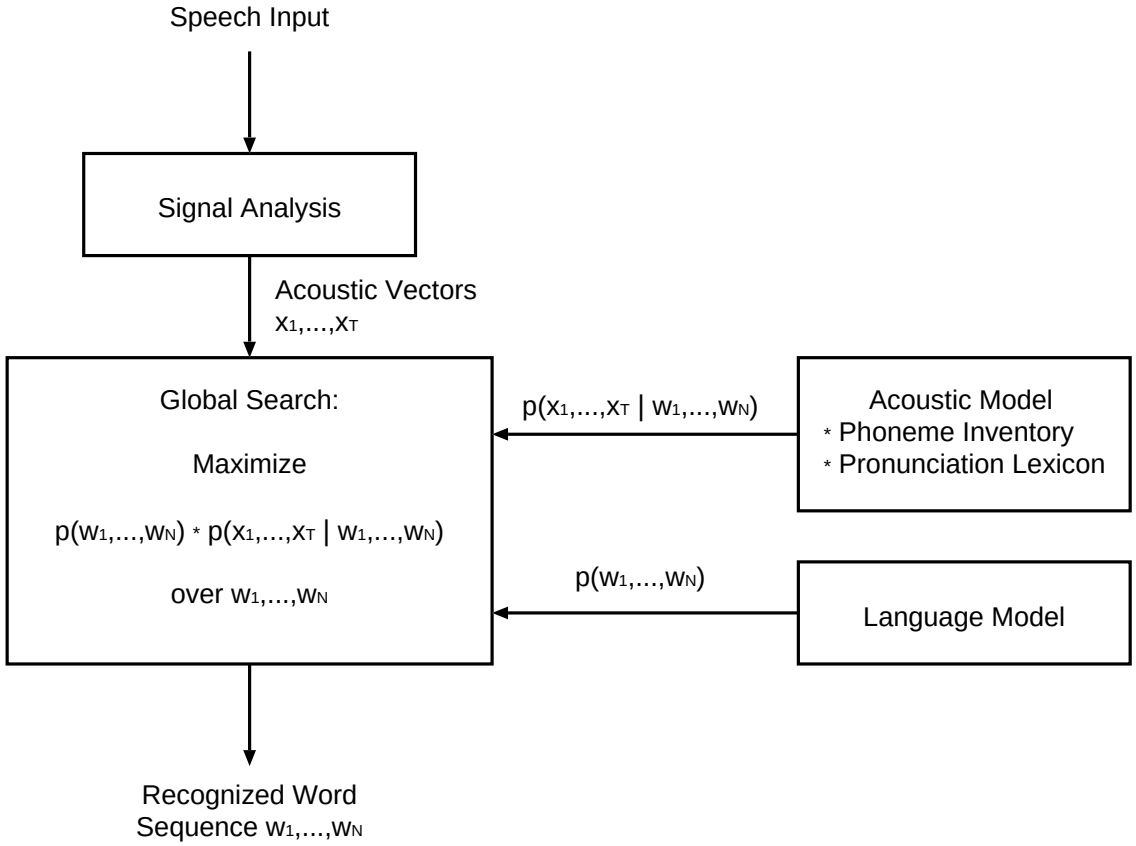

Figure 1.1: Basic architecture of a statistical automatic speech recognition system. [Ney 90].

1.2 Signal Analysis

The signal analysis module aims at providing the speech recognition system with a sequence of acoustic vectors. The acoustic vectors build a parameterization of the speech waveform observed at the microphone. The signal analysis should remove as much information irrelevant for the speech recognition process as possible, for instance intensity and pitch, and retain only the information relevant for the *content* of the utterance. Thus the acoustic vectors should fulfill the following requirements:

- be of low dimensionality to allow a reliable estimation of the free parameters of the speech recognition system,

- be independent of the speaker and recording environment, i.e. only dependent on the contents of the spoken word sequence,
- be characteristic for the sub-word unit to allow an optimal discrimination between the different acoustic models.

The signal analysis of today's state-of-the-art speech recognition systems is based on a short term spectral analysis [Rabiner & Schafer 78], usually a Fourier analysis. Two procedures for further processing and smoothing are widely used: *Mel frequency Cepstral coefficients* (MFCC) [Davis & Mermelstein 80] and *perceptual linear prediction* (PLP)[Hermansky 90]. In this work the MFCC-based signal analysis front-end of the RWTH speech recognition is used. The basic steps of the RWTH signal analysis front-end are shown in Figure 1.2 and described in detail in [Welling 99].

After some preprocessing like preemphasis and windowing, the Fourier power spectrum of the speech waveform is computed for each time frame with a frame shift of 10ms and a window length of 25ms. The frequency axis of this power spectrum is warped according to the *Mel frequency* scale to adjust the spectral resolution to that of the human ear [Young 93]. Afterwards, a filter bank is applied and the logarithm is taken to reduce the dynamic range and a Cepstrum transformation is applied to the log filter bank coefficients to remove the correlation between the different outputs. The dimensionality of the Cepstral vector is reduced by omitting the highest Cepstral coefficients for smoothing. Mean, variance and energy normalization are applied subsequently to the MFCCs to reduce a potential mismatch introduced by different transmission channels and recording environments. An alternative MFCC computation scheme which integrates the frequency axis warping into the Cepstrum transform and computes the Cepstral coefficients directly on the log-power spectrum will be presented in Chapter 7.

A commonly used method to include dynamic information is to augment the original vector with the first and second derivatives yielding a high dimensional vector. A more general approach is *linear discriminant analysis* (LDA) [Fisher 36, Duda & Hart$^+$ 01]. The LDA is a linear transformation which projects a feature space into a lower dimensional subspace such that the class separability for distributions with equal variances is maximized. In the RWTH system several successive feature vectors are augmented (either three vectors containing derivatives or seven vectors without derivatives, depending on the specific task). This high dimensional feature space is reduced using LDA to a dimension of typically 25 (telephone data) or 33 (broadband data). It will be shown in Chapter 5 that LDA fits into a more general framework of linear transformation schemes.

Especially the demand of speaker independence on the acoustic vectors is hard to meet. The above mentioned MFCC and PLP features are for instance also used for speaker identification tasks [Doddington & Przybocki$^+$ 00], which means that there is still a lot of information of the given speaker contained in those features. Several methods have been developed to cope with the speaker dependence of the

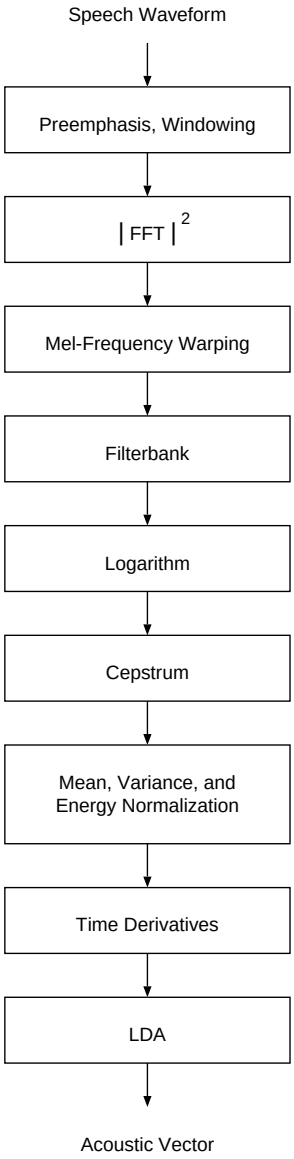

Figure 1.2: Scheme of the RWTH signal analysis front end.

acoustic feature vectors: *speaker normalization*, which tries to reduce the speaker dependency by transforming the acoustic feature vectors, and *speaker adaptation*, which tries to adjust the model parameters of the speech recognition system to the characteristics of the given speaker[1].

[1]This distinction is not orthogonal as some normalization schemes can be formulated as adaption and vice versa, c.f. Chapters 4 and 5

1.3 Acoustic Modeling

The aim of acoustic modeling is to provide a stochastic model $p(x_1^T|w_1^N)$ for the realization of a sequence of acoustic vectors x_1^T given a word sequence w_1^N. The acoustic model is a concatenation, according to a pronunciation lexicon, of the acoustic models for the basic sub-word units that the speech recognition system utilizes.

Dependent on the amount of training data and the desired model complexity, the sub-word units are usually build of whole words, syllables, phonemes or phonemes in context. Smaller units than words enable the speech recognition system to recognize words which have not been seen in the training data and to ensure that enough instances of each unit have been observed in training to allow a reliable parameter estimation. In large vocabulary speech recognition (LVCSR) the most commonly used sub-word units are phonemes in a context of one or two adjacent phonemes, so called *triphones* and *quinphones*, respectively. Context-dependent phonemes (*n-phones*) are used to care for the different pronunciations of a phoneme depending on the surrounding phonemes.

The acoustic realizations of a sub-word unit differ significantly with the speaking rate. To model the variations in speaking rate, *Hidden Markov Models (HMM)* have been established as de-facto standard for speech recognition systems [Baker 75, Rabiner 89]. An HMM is a stochastic finite state automaton consisting of a number of states and transitions between the states. The probability $p(x_1^T|w_1^N)$ is extended by an unobservable (hidden) random variable representing the states:

$$p(x_1^T|w_1^N) = \sum_{s_1^T:w_1^N} p(x_1^T, s_1^T|w_1^N)$$

where the sum is taken over all possible state sequences s_1^T for a given word sequence w_1^N. Using Bayes' identity this can be rewritten as

$$p(x_1^T|w_1^N) = \sum_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|x_1^{t-1}, s_1^t; w_1^N) \cdot p(s_t|x_1^{t-1}, s_1^{t-1}; w_1^N) \ .$$

This equation can be further simplified by applying a first order Markov assumption [Duda & Hart$^+$ 01]. The probabilities $p(x_t|x_1^{t-1}, s_1^t; w_1^N)$ and $p(s_t|x_1^{t-1}, s_1^{t-1}; w_1^N)$ are assumed not to depend on previous observations but only on the states (Markov) and on the immediate predecessor state only (first order):

$$p(x_1^T|w_1^N) = \sum_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t; w_1^N) \cdot p(s_t|s_{t-1}; w_1^N) \ . \tag{1.2}$$

Thus the probability $p(x_1^T|w_1^N)$ is split into the acoustic *emission probability* $p(x_t|s_t; w_1^N)$, denoting the probability to observe an acoustic vector x_t while being in state s_t, and the *transition probability* $p(s_t|s_{t-1}; w_1^N)$ for a transition from

state s_{t-1} to state s_t. Often the sum in Eq. (1.2) is approximated by the maximum; this approximation is usually called *Viterbi* or *Maximum* approximation [Ney 90]

$$p(x_1^T|w_1^N) \approx \max_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t;w_1^N) \cdot p(s_t|s_{t-1};w_1^N) \ . \tag{1.3}$$

The Eqs. (1.2) and (1.3) can be solved efficiently using the *forward-backward* algorithm [Baum 72, Rabiner & Juang 86] or *dynamic programming* [Bellman 57, Viterbi 67, Ney 84].

An example of an HMM for a part of the word "seven" is given in Figure 1.3. The topology used in this work has been introduced by Bakis [Bakis 76]: the basic HMM consists of six subsequent states where each two successive states are identical. Between the state transitions to the same state (*loop*), the next state (*forward*), and the next state but one (*skip*) are allowed. Using a frame-shift of 10ms (cf. Section 1.2) the path through the HMM with forward transitions only amounts to 60ms, which is close to the average duration of phonemes for most languages. This *6-state* HMM has a minimum duration of 30ms (only skip transitions). This has been found to be too long for fast conversational speech, e.g. on the Verbmobil II corpus [Molau 03]. In this case a *3-state* model is used where the two identical states are merged into a single one, which reduces the minimum length of the HMM.

The emission probabilities $p(x_t|s_t;w_1^N)$ of an HMM can be modeled by discrete probabilities [Jelinek 76], semi-continuous probabilities [Huang & Jack 89] or as continuous probability distributions [Levinson & Rabiner$^+$ 83]. A commonly used model for continuous probability distributions are mixture densities made up of a weighted sum of either Gaussian or Laplacian probability densities; a systematic comparison of Gaussian or Laplacian probability density functions can be found in [Chen & Eide$^+$ 99]. For Gaussian mixture densities, which are used in the RWTH system, the emission probabilities are given as follows:

$$p(x|s;w_1^N) = \sum_{l=1}^{L_s} c_{sl}\, \mathcal{N}(x|\mu_{sl},\mathbf{\Sigma};w_1^N) \tag{1.4}$$

where c_{sl} denotes the mixture weights with the constraint $\sum_{l=1}^{L_s} c_{sl} = 1$ and $\mathcal{N}(x|\mu,\mathbf{\Sigma})$ denotes the normal distribution with mean μ and covariance $\mathbf{\Sigma}$. In the RWTH system the covariances $\mathbf{\Sigma}$ are tied over all states[2] and are modeled by a diagonal matrix to care for data sparseness problems and for efficiency reasons. The set of parameters $\theta = \{\{\mu_{sl}\},\{c_{sl}\},\mathbf{\Sigma}\}$ is estimated using *Maximum Likelihood* estimation in combination with the *Expectation Maximization* algorithm [Dempster & Laird$^+$ 77].

When using n-phones as basic sub-word units the number of states to be modeled raises exponentially with the context length. Thus a large number of n-phones will

[2] i.e. all states share the same variance

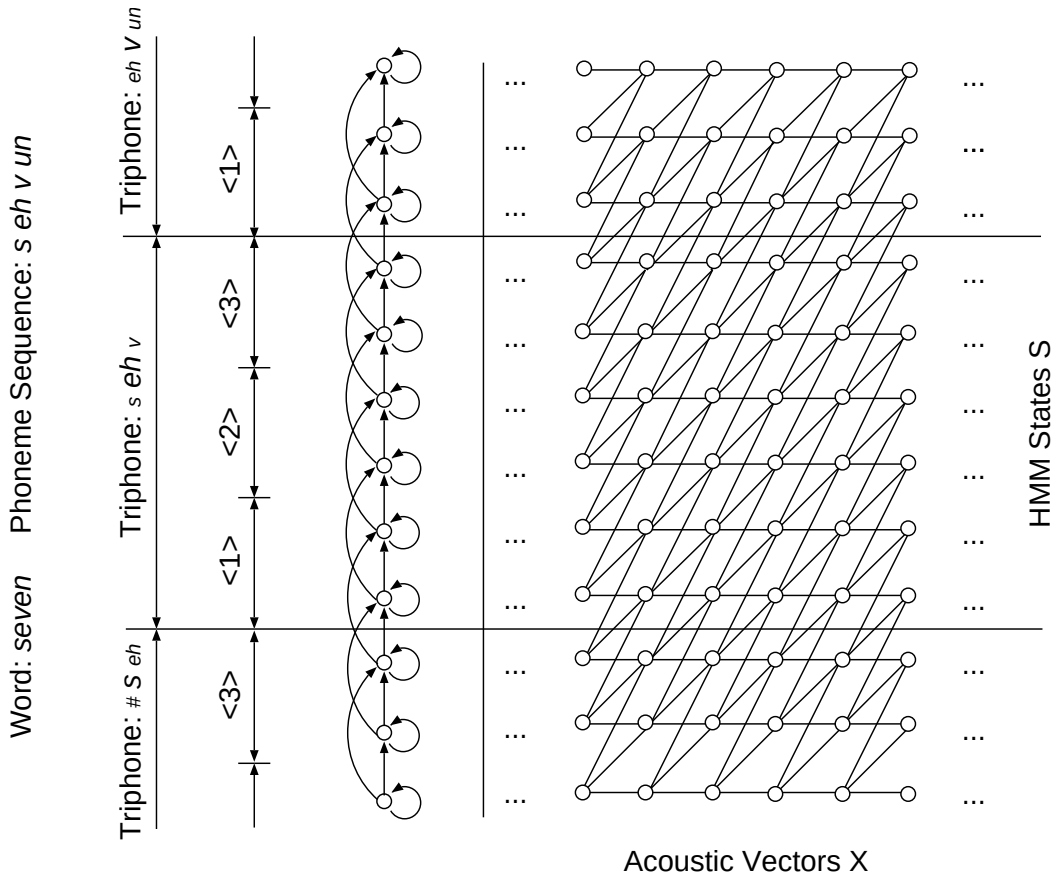

Figure 1.3: 6-state hidden Markov model in Bakis topology for the triphone ${}_{s}eh_{v}$ in the word "seven". The HMM segments are denoted by <1>, <2>, and <3>.

have no or too few observations for a reliable parameter estimation. Therefore several states are tied together yielding *generalized* n-phone models [Young 92]. Decision tree based state clustering is used in almost all LVCSR systems. The main advantage of this top-down clustering method is that no back-off models have to be trained and unseen n-phones will be assigned to an appropriate HMM state. Details of the state clustering in the RWTH system can be found in [Beulen & Ortmanns$^+$ 99]. As the pronunciation of a phoneme depends on the surrounding phonemes, a phoneme at a word boundary is pronounced differently dependent on the predecessor and successor words. This coarticulation effect is modeled explicitly using *across-word* n-phones [Hon & Lee 91, Odell & Valtchev$^+$ 94], which take into account the ending and beginning phonemes of the adjacent words as left and right context, respectively. Details of the across-word model implementation for the RWTH system can be found in [Sixtus 03].

1.4 Language Modeling

The language model $p(w_1^N)$ provides an a-priori probability for a word sequence $w_1^N = w_1, \ldots, w_N$. The syntax, semantics and pragmatics of the language to be recognized are implicitly covered by this statistical model. Due to the unlimited number of possible word sequences further model assumptions have to be applied in order to estimate a reliable model. For LVCSR *m-gram* language models [Bahl & Jelinek$^+$ 83] have become widely accepted. m-gram language models assume that the word sequence follows an $(m-1)$-th order Markov process: the probability of the word w_n depends on the $(m-1)$ predecessor words only. Thus the probability $p(w_1^N)$ can be expressed as

$$\begin{aligned} p(w_1^N) &= \prod_{n=1}^{N} p(w_n|w_1^{n-1}) \\ &\underset{\text{model assumption}}{=} \prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1}) \; . \end{aligned} \tag{1.5}$$

The word sequence $h_n = w_{n-m+1}^{n-1}$ is denoted as *history* of length m of the word w_n with the definitions $h := w_1^{n-1}$ if $n < m$ and $h := \emptyset$ if $n - 1 < n - m + 1$, e.g. at the boundary $p(w_1|w_1^0) = p(w_1)$.

A commonly used measure for the evaluation of language models is the *perplexity* PP

$$PP = \left[\prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1}) \right]^{-1/N} \; .$$

The log-perplexity is equal to the entropy of the model and can be interpreted as average number of choices to continue a word sequence w_{n-m+1}^{n-1} at position n. When using the perplexity as optimization criterion for training the language model, closed form solutions for $p(w|h)$ can be derived which are equal to the relative frequency of the word sequence on the training corpus. The number of possible m-grams increases exponentially with the history length m. Thus, for a large vocabulary W, a considerable amount of m-grams will be unseen in training or have too few observations for a reliable estimation of $p(w|h)$, even for very large training corpora. Therefore smoothing methods have to be applied. The smoothing is based on *discounting* in combination with *backing-off* or *interpolation* [Katz 87, Ney & Essen$^+$ 94, Generet & Ney$^+$ 95, Ney & Martin$^+$ 97]. Discounting subtracts probability mass from seen events which is than distributed over all unseen events (backing-off) or over all events (interpolation), usually in combination with a language model with shorter history. The parameters of the smoothed language model can be estimated using a cross-validation scheme like *leaving-one-out* [Ney & Essen$^+$ 94]. Details of the language model implementation in the RWTH system can be found in [Wessel & Ortmanns$^+$ 97].

1.5 Search

The search module of the speech recognition system combines the two knowledge sources acoustic model and language model as depicted in Fig. 1.1. The objective of the search is to find that word sequence which maximizes the a-posteriori probability for a given sequence x_1^T of acoustic feature vectors according to Eq. (1.1)

$$\begin{aligned}[w_1^N]_{\text{opt}} &= \underset{w_1^N}{\text{argmax}}\, p(w_1^N|x_1^T) \\ &= \underset{w_1^N}{\text{argmax}} \left\{ p(w_1^N) \cdot p(x_1^N|w_1^N) \right\} \ .\end{aligned}$$

If the language model is given by an m-gram model (Eq. (1.5)) and the acoustic model is an HMM as given in Eq. (1.2), the following optimization problem has to be solved by the search module:

$$\begin{aligned}[w_1^N]_{\text{opt}} &= \underset{w_1^N}{\text{argmax}} \left\{ \left[\prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1}) \right] \cdot \left[\sum_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t;w_1^N) \cdot p(s_t|s_{t-1};w_1^N) \right] \right\} \\ &\cong \underset{w_1^N}{\text{argmax}} \left\{ \left[\prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1}) \right] \cdot \left[\max_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t;w_1^N) \cdot p(s_t|s_{t-1};w_1^N) \right] \right\} \ .\end{aligned} \tag{1.6}$$

In the second step the Viterbi approximation has been applied to the HMM, which significantly reduces the complexity of the optimization problem. Eq. (1.6) can be solved efficiently using *dynamic programming* [Bellman 57]. Dynamic programming exploits the mathematical structure and divides the problem in subinstances. As in all search problems, the search can be organized in two different ways: a *depth-first* and *breadth-first* search. The depth-first strategy is used by the A^*-*search* or *stack-decoding* algorithm. Here the state hypotheses are expanded time-asynchronously dependent on a heuristic estimate of the cost to complete the path [Jelinek 69, Paul 91]. The performance of the A^*-search relies strongly on the quality of this estimate; the convergence to a global optimum is guaranteed if the estimate is a lower bound of the true costs. Additionally, the search space is minimal if the estimate is equal to the true costs.
The breadth-first search design is used by the *Viterbi* search where all state hypotheses are expanded time-synchronously [Vintsyuk 71, Baker 75, Sakoe 79, Ney 84]. In this approach the probabilities of all hypotheses up to a given time frame are computed and thus can be compared to each other. This allows to reduce the search space significantly by pruning unlikely hypotheses early in the search process. Especially in the breadth-first approach an efficient pruning is necessary as the number of possible word sequences with maximum length N grows exponentially with N. Thus a full optimization of Eq. (1.6) is only feasible for small vocabulary sizes $|W|$. For

large vocabulary sizes approximations have to be made. Instead of finding the exact optimal solution of Eq. (1.6) the goal is changed to find a sufficiently good solution with much less effort. In the so-called *beam-search*, only that fraction of the hypotheses are expanded whose likelihood is sufficiently close to that of the best hypothesis of the given time frame [Lowerre 76, Ney & Mergel$^+$ 87, Ortmanns & Ney 95]. Beam-search does not guarantee to find the globally best word sequence. This optimal sequence may have been pruned at an intermediate search stage due to a poor likelihood. However, if the pruning parameters are adjusted properly no significant search errors occur and the search effort is reduced considerably.

Several other methods can be applied to further reduce the computational complexity of the Viterbi or beam-search:

- *Lexical prefix tree* [Ney & Haeb-Umbach$^+$ 92]: the pronunciation lexicon is organized as lexical prefix tree. A considerable amount of search effort is spend on the first few phoneme-models of words that will be pruned away later in the search process. Since the word identity is not known at the start of a hypothesis any longer, this organization causes some overhead in computational effort. Thus a lexical prefix tree is only beneficial for beam search and large vocabulary sizes.

- *Look-ahead*: these techniques try to incorporate approximated information of future search stages. The *language model look ahead* is used to overcome some of the overhead caused by the tree organization of the lexicon [Steinbiss & Tran$^+$ 94, Odell & Valtchev$^+$ 94, Alleva & Huang$^+$ 96, Ortmanns & Ney$^+$ 96a]: At each node of the tree the probability of the most likely word reachable from that node is taken into account to make the language model pruning more effective. The *phoneme look ahead* estimates the acoustic probability of a few future time frames using a simplified acoustic model (e.g. monophone models instead of triphone models) and thus enhances the acoustic pruning [Ney & Haeb-Umbach$^+$ 92, Haeb-Umbach & Ney 94, Ortmanns & Ney$^+$ 96b].

- *Fast likelihood computation:* a considerable amount of computation time of an high-performance speech recognition system is spend on computing the acoustic emission probabilities (cf. Eq. (1.4)). Typical high-performance speech recognition systems make use of several 100.000 densities. Proposed approaches to reduce the effort of likelihood computation are based on structuring the search space [Ramasubramansian & Paliwal 92, Fritsch 97], quantization of the feature vectors [Bocchieri 93, Ortmanns & Ney$^+$ 97b] or by partitioning the feature space [Nene & Nayar 96]. Details of the fast likelihood computation used in the RWTH system can be found in [Ortmanns 98]. Additionally, utilizing the SIMD[3] instructions of modern CPUs for a par-

[3]single instruction, multiple data

allelized likelihood computation significantly reduces the computation time [Kanthak & Schütz$^+$ 00].

Multi-pass search strategies are used either to enhance the accuracy of the speech recognition system or to speed up the system without significant loss in accuracy. In both cases, a preliminary recognition pass is performed and the results are stored as *N-best lists* or *word graphs* for further processing. In *N*-best lists, the N most likely word sequences are stored [Schwartz & Chow 90, Schwartz & Austin 91, Ney & Oerder 93]. A word graph is a directed acyclic graph whose arcs contain the word labels of word sequence alternatives [Schwartz & Austin 91, Ney & Aubert 94, Ortmanns & Ney$^+$ 97a]. In systems optimized for accuracy the first pass is performed with fully trained acoustic models gaining high-accuracy word-graphs. All subsequent passes than restrict the search space to those hypotheses contained in the word graph. The second pass utilizes more complex acoustic models with a longer context and/or more complex language models like 7-gram models. Additionally, speaker dependent transformations are estimated on the results of the first pass to adapt the speech recognition system to the given speaker. Modern systems make use of five or more passes to obtain the final result [Schwartz & Colthurst$^+$ 04, Evermann & Chan$^+$ 04].
The approach for systems optimized for runtime speed is a bit different. Here a very fast first pass is carried out with simplified acoustic and/or language models, in order to rapidly produce a preliminary transcription or a large word graph for further processing [Saon & Zweig$^+$ 03]. Afterwards, a second, refined acoustic model is adapted in several steps using the recognition results of the first pass. In subsequent passes, the final output is then obtained using these adapted acoustic models.

Chapter 2

State of the Art in Speech Recognition

Linear transformations have a long history in speech recognition research. This Chapter gives an overview of the most important approaches related to this thesis.

2.1 Feature Space Transformations

One widely used group of feature space transformations are feature space projections. Feature space projections aim at reducing the dimensionality of the acoustic feature space in order to find an optimal subspace for discrimination. After the transformation, the dimensions of the feature space are sorted according to their importance for class discrimination. The dimensionality of the feature space is then reduced by omitting those dimensions which contain little class information.

The most common feature space projection is *linear discriminant analysis (LDA)* [Fisher 36]. It was first introduced for discrimination of syllables by [Hunt 79]. The basic concepts of using LDA for speech recognition were presented in [Brown 87]. LDA has been applied for continuous speech recognition for the first time in [Doddington 89]. A successful application of LDA for large vocabulary has been reported in [Haeb-Umbach & Ney 92].

The approach of LDA has been extended in [Kumar & Andreou 98] to overcome the constraint of equal covariances of the class distributions, called *heteroscedastic discriminant analysis (HDA)*. Additionally, the authors present a general theory for linear feature space transformations including dimensionality reducing transformations in a maximum-likelihood framework. It was also shown that the LDA can be derived from a maximum-likelihood approach and fits well in this general framework. The authors achieved an improvement on the small vocabulary TI-DIGITS task from 0.67% word error rate with LDA to 0.59% word error rate with HDA using full covariance matrices and from 2.29% word error rate (LDA) to 1.65% word error rate (HDA) with diagonal covariance matrices.

Based on that work, both approaches presented in [Gales 97, Gales 99] and [Gopinath 98] applied an HDA-like transformation, but without dimensionality re-

duction. In [Gales 97, Gales 99] the approach was called *semi-tied covariances*, in [Gopinath 98] *maximum-likelihood linear transformation (MLLT)*. The idea of this approach is to transform the feature space such that the resulting covariance matrices are as diagonal dominant as possible. In [Gales 99] the author reports consistent improvements on the 1994 ARPA Hub 1 test set of about 12% rel. In [Gopinath 98] the same transformation was tested on the 1996 DARPA Hub 4 test set. The author reports improvements of 8.5% rel. on the F0 subset (planned speech) and of 2.5% rel. on the F1 subset (spontaneous speech).

The semi-tied covariances or MLLT approach has been extended in [Olsen & Gopinath 02] by increasing the degrees of freedom of the transformation, called *extended maximum-likelihood linear transformation (EMLLT)*. The inverse covariance matrices are taken from a subspace of a chosen number of rank one matrices. The complexity of the (inverse) covariance modeling can be adjusted from diagonal covariances (which is equal to MLLT) up to full covariance modeling in a consistent way by choosing the number of rank one matrices which build up the subspace to which the inverse covariances are restricted. The authors report recognition test results on an inhouse car navigation task, the recognition performance could be improved by 9.5% rel. using MLLT and by 35% rel. using EMLLT. A further extension to EMLLT has been presented in [Axelrod & Olsen 02]. Instead of using rank one matrices, the subspace is spanned by a chosen number of arbitrary symmetric matrices. Although the overall recognition accuracy could not be improved compared to full covariance modeling, the authors achieved consistently better results compared to EMLLT and diagonal covariance modeling given equal number of parameters. Thus this approach allows for a significant reduction in model complexity (and thus memory requirements and computation time) with only little loss in recognition accuracy.

2.2 Linear Speaker Adaptation Transformations

One of the first approaches to apply linear transformations for speaker adaptation was presented in [Jaschul 82] by means of spectral mapping. In [Choukri & Chollet$^+$ 86] a method was presented based on canonical correlation analysis. Here, both the reference data as well as the test data from the new speaker are transformed into a common feature space by estimating two different transformations. These spectral mapping approaches have been applied in an HMM framework by [Class & Kaltenmeier$^+$ 90] for an isolated word task.

An approach which is used in virtually all state-of-the-art speech recognition systems is called *Maximum Likelihood Linear Regression (MLLR)* and was proposed in [Leggetter & Woodland 95a]. An affine transformation is applied to the mean vectors of the acoustic emission probabilities of the HMM. Usually, several HMM states are grouped into adaptation classes which share the same transformation ma-

trix. The adaptation classes are defined manually based on broad phonetic classes or data driven by clustering adjacent mean vectors of the HMM emission probability. In [Leggetter & Woodland 95a] the authors used a data driven approach with a predetermined number of adaptation classes. The authors report improvements of 21% rel. using one adaptation class and achieved the best result with an improvement of 37% rel. using 15 adaptation classes on the medium vocabulary ARPA Resource Management task. The use of regression class trees for the definition of the adaptation classes has been proposed in [Leggetter & Woodland 95b]. The recognition performance could be improved by 20% rel. using a regression class trees compared to one global adaptation matrix. A comparison with the data driven approach used in [Leggetter & Woodland 95a] was not given. A detailed comparison of regression class trees is reported in [Haeb-Umbach 01]. If the number of adaptation data is small, the estimation of the regression matrix is unstable and the regression matrix tends to be singular [Leggetter & Woodland 95a]. A possible solution is the introduction of a threshold for a minimum number of adaptation utterances. In [Neumeyer & Sankar$^+$ 95] a block-diagonal structure of the transformation matrix was suggested, which decreases the number of parameters to be estimated. The authors achieved an improvement on the WSJ Spoke 3 test set of 8% rel. compared to a full transform and 15% rel. compared to a diagonal transform.

The transformations presented so far only affected the mean vectors of the probability distribution. An approach to also transform the variances has been presented in [Gales & Woodland 96]. The variances of the Gaussian probability distributions were transformed using a second linear transformation. Using a diagonal transformation, the authors achieved improvements on different corpora ranging from 1.5% to 4.3% relative reduction in word error rate when adapting mean and variances compared to adapting the means only. These additional improvements of adapting the variances in addition to the means are quite small compared to the improvements of adapting the mean vectors only, which range from 9.8% to 13.1% relative reduction in word error rate compared to the baseline with no adaptation.

Another way of adapting the variances is to use the same transformation as for the mean vectors. This has been proposed in [Digalakis & Rtischev$^+$ 95] for diagonal transformations and later in [Gales 98] for the general case. This approach is named *constrained MLLR, (C-MLLR)*; the term constrained is used because the transformation matrices for the mean vectors and the variances are forced to be equal in contrast to the usual (unconstrained) MLLR where the transformation matrices for the mean vectors and variances are independent of each other. The C-MLLR approach is equivalent to transforming the feature vectors instead of the model parameters, which allows an efficient implementation for the case of using one global transformation only, i.e. no regression classes. The comparison of C-MLLR with the MLLR approach presented in [Gales & Woodland 96] revealed little differences in terms of the recognition accuracy.

The adaptation schemes presented above were applied to given model parameters

in test only, the training part was unaltered. An approach to use linear transformations during the training of the speech recognition system has been presented in [Anastasakos & McDonough+ 96] and is called *speaker adaptive training*. This approach estimates the model parameters together with appropriate MLLR transformations for the training speakers. The authors achieved improvements of 8.9% and 11.4% relative reduction in word error rate for the 20k WSJ H1 and the 5k WSJ S0 test sets, respectively. Despite this improvements, a major drawback of speaker adaptive training is a substantial increase in complexity in terms of time and memory requirement and therefore speaker adaptive training is not as common as MLLR adaptation in state-of-the-art speech recognition systems. However, in [Gales 98] the author applied the C-MLLR approach also in training. As C-MLLR can be formulated as transformation of the feature vectors, the estimation formula for the model parameters remain almost unchanged. The gains obtained by speaker adaptive training with C-MLLR are similar to those obtained by MLLR given above.

One important problem when applying speaker adaptation in an unsupervised mode are the recognition errors of the first pass. This becomes even more important on tasks with a high word error rate, for instance automatic recognition of conversational speech. Confidence measures have been investigated to aid the adaptation process dealing with faulty transcripts [Zeppenfeld & Finke+ 97, Anastasakos & Balakrishnan 98, Nguyen & Gelin+ 99, Wallhoff & Willet+ 00]. In all these works the confidence measures have been used to mark possible recognition errors and to exclude those words from the adaptation process. The results reported are inconsistent and rely very much on the specific confidence measure ranging from only marginal improvements [Anastasakos & Balakrishnan 98, Nguyen & Gelin+ 99] to 3.3% [Zeppenfeld & Finke+ 97] and 4.1% [Wallhoff & Willet+ 00] relative improvement in word error rate, respectively.

A major drawback of using confidence measures in the manner described above is the reduction of adaptation data. Therefore a word graph has been used to collect the statistics needed for the adaptation process [Padmanabhan & Saon+ 00, Uebel & Woodland 01]. This approach uses a weighted sum of the alternative hypotheses represented in the word graph rather than discarding parts with low confidence. While in both works the difference of confidence-based and lattice-base MLLR has been shown to be marginal if only one adaptation pass is applied, in [Uebel & Woodland 01] it has been shown that lattice-based MLLR gives a relative improvement in word error rate of about 3% if used in several successive recognition and adaptation iterations.

Another major problem in speaker adaptation is data sparseness, especially for fast speaker adaptation. To estimate the transformation matrix reliably, several minutes of speech data are needed. Although MLLR adaptation can be applied iteratively, the performance is far worse than in a two-pass mode. This problem has been addressed using prior distributions for the MLLR matrix [Chesta & Siohan+ 99,

Chou 99]. The authors achieved a relative improvement in word error rate over the conventional MLLR technique up to 13.5% for limited adaptation data and ended up with equal performance for the case of many adaptation data.

2.3 Vocal Tract Normalization

The idea of scaling the frequency axis of the speech signal to deal with gender specific variations has been proposed first in [Wakita 77] for isolated vowel recognition. The idea has been picked up later in [Acero 90, Acero & Stern 91] for small vocabulary speech recognition with an improvement of 10% relative reduction in word error rate using a bilinear transformation function. For large vocabulary, vocal tract normalization (VTN) has been proposed in [Eide & Gish 96, Lee & Rose 96, Wegmann & McAllaster$^+$ 96]. In [Eide & Gish 96], several warping functions have been compared showing little differences in recognition performance. The warping factor has been estimated using the median position of the third formant. The improvements range from 8% to 6% relative reduction in word error rate for 5 hours and 63 hours of training data, respectively. A maximum likelihood estimation of the warping factors has been suggested in [Lee & Rose 96] together with an iterative scheme to estimate the warping factors on the training speakers. The estimation of the warping factors in test is either done by a forced alignment using a preliminary recognition pass or is based on a text-independent Gaussian mixture model without the need of a first recognition pass. The recognition performance could be improved by 15% relative reduction in word error rate using the one-pass approach and by 20% using a preliminary recognition pass. A similar approach has been presented in [Wegmann & McAllaster$^+$ 96], where a piece-wise linear warping function has been applied yielding improvements of 12% relative reduction in word error rate.

The transformation functions used by then were either convex or concave. In [McDonough 98] all-pass transforms have been proposed as transformation function, which are an extension to bilinear functions. The word error rate could be lowered by 7% relative for the bilinear transformation and by 8% for the more flexible all-pass transform. Additionally, it was shown that VTN amounts to a linear transformation in the Cepstrum domain when using the all-pass transform. This was previously shown in [Acero 90, p.119] for the case of a bilinear transformation.

The relationship between VTN and linear transformations, especially MLLR, has been investigated in [Pye & Woodland 97], showing that improvements obtained by unconstrained MLLR and VTN are largely additive. However, in [Uebel & Woodland 99] it has been shown that for the case of C-MLLR there is no additional improvement from VTN, indicating that both approaches may not be independent of each other.

Several approaches have been proposed to realize VTN as a linear transformation of MFCC. In [Cox 00], the author used a tri-diagonal transformation matrix

to transform the MFCC features with the restriction of all elements on one (secondary) diagonal being equal, thus the matrix consists of only three free parameters. Phoneme recognition test results with a supervised adaptation yielded only moderate improvements. The linearity of VTN for the case of a bilinear transformation function was utilized in [Emori & Shinoda 01], where the authors calculated the transformation matrix approximatively for small values of the warping parameter. The computational cost of estimating the warping factor could be lowered while maintaining the recognition performance.

Based on considerations of the effect of frequency warping on the filter bank, band-diagonal matrices have been proposed to enhance MLLR adaptation with limited adaptation data [Uebel & Woodland 99, Afify & Siohan 00]. In [Afify & Siohan 00], the authors have proposed to constrain the MLLR matrix to a band-diagonal matrix, a structure which had previously been been observed in [Uebel & Woodland 99]. Applying this restriction to the MLLR matrix, the recognition performance could be improved significantly on the non-native Spoke3 test set of the Wall Street Journal task for various amounts of adaptation data and different numbers of bands [Afify & Siohan 00]. The same approach has later been repeated on a Chinese isolated word task with similar results in [Ding & Zhu$^+$ 02].

2.4 Summary

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

- Linear transformations have been shown to improve recognition accuracy at various stages of the recognition process.

- Feature space transformations have been investigated extensively to define subspaces for the acoustic models, resulting in a wide variety of approaches which are often closely related or even equivalent (for instance semi-tied covariances and MLLT).

- Linear transformations have been proven very successful for speaker adaptation and have become a quasi-standard in up-to-date speech recognition systems.

- In connection with the problem of recognition errors in an unsupervised two-pass adaptation mode, confidence measures and lattice-based adaptation have been studied with inconsistent results.

- Vocal tract normalization provides an effective and online capable method to deal with speaker-specific variations.

- Several efforts have been made to represent vocal tract normalization as linear transformation in the Cepstral domain, but no exact interrelationship could be derived.
- No clear picture has emerged on the use of C-MLLR vs. MLLR.
- There is some evidence that C-MLLR and vocal tract normalization are not independent of each other.

Chapter 3

Scientific Goals

Speaker adaptation and normalization techniques play an important role in current state-of-the-art automatic speech recognition systems. Although widely used, no clear picture has emerged from literature on the connection of the various approaches. In particular, the interconnection of vocal tract normalization and linear transformations has been addressed but has not been understood sufficiently.

The following topics will be studied in this thesis:

1. Although speaker adaptation and normalization techniques are widely used, no consistent framework has been developed. In fact, speaker adaptation and normalization are oftenly considered as different approaches to reduce the speaker specific variability in the acoustic signal. In this thesis a common mathematical framework to describe both speaker adaptation and normalization will be developed in the context of Bayes' decision rule. Moreover, it will be proven that from the conceptual point of view both techniques are equivalent if Gaussian mixture distributions are used as emission probabilities of the hidden Markov model.

   Using this common framework, a unified view of most linear transformations used in speech recognition will be discussed and similarities and differences of the approaches will be identified and analyzed.

2. The standard Mel frequency Cepstral coefficients (MFCC) features make use of a twofold smoothing: one is provided by the filter bank and the other by omitting higher coefficients of the subsequent Cepstrum transformation. This twofold smoothing is disadvantageous because the optimal numbers of filter banks and Cepstral coefficients have to be optimized again for each new application. A new approach will be introduced that does not make use of a filter bank and the smoothing is provided solely by the Cepstrum transformation.

3. Using this enhanced MFCC computation scheme, it will be derived analytically that vocal tract normalization can always be expressed as a linear transformation of the Cepstral coefficients for *arbitrary* invertible warping functions. The transformation matrix will be explicitly calculated for three common warping

functions. Based on some general characteristics of typical warping functions, the shape of the matrices for different warping functions will be investigated.

4. Based on the structure of the transformation matrix for vocal tract normalization, a restriction of the maximum likelihood linear regression (MLLR) matrix will be investigated, which allows a more robust estimation for the case of limited adaptation data.

5. By representing VTN as linear transformation, the Jacobian determinant can now, for the first time, be taken into account for any warping function. So far, the Jacobian determinant has been simply omitted or taken into account only for very special warping functions. In this work the influence of the Jacobian determinant on the warping factor estimation and thus the performance of vocal tract normalization will be studied exemplary for a piece-wise linear warping function.

6. Word posterior probabilities have been proven to provide a good confidence measure. It will be shown that using word posterior probabilities as a confidence measure can significantly help to solve the problem of recognition errors in a two-pass adaptation scheme.

7. Regression classes are commonly used to enhance the MLLR adaptation performance. However, the use of regression classes needs a substantial amount of adaptation data to gain improvements. In this thesis several approaches will be presented to benefit from regression classes even in the case of limited adaptation data.

The remainder of this thesis is organized as follows: in Chapter 4 a common framework of adaptive acoustic modeling will be developed. Afterwards, a unified view of linear transformations used in speech recognition will be given in Chapter 5. Based on a modified signal analysis, which will be described in Chapter 6, it will be shown in Chapter 7 that vocal tract normalization amounts to a linear transformation in the Cepstral space. Due to this linearity, the transformation matrix and thus the Jacobian determinant can be calculated. The effect of incorporating the Jacobian determinant for vocal tract normalization will be discussed in Chapter 8. In Chapter 9, several improvements of the MLLR adaptation will be presented, in particular a refined modeling of adaptation classes, the use of confidence measures for enhanced MLLR adaptation and a restriction of the MLLR matrix based on the results of Chapter 7. This thesis will be concluded by a summary of the scientific contributions in Chapter 10 and an outlook in Chapter 11. Details on some more complex calculations as well as details on the speech corpora and recognition settings used in the experimental evaluations will be given in the Appendix.

Chapter 4

Adaptive Acoustic Modeling

4.1 Introduction

A speaker independent automatic speech recognition (SI-ASR) system has to cope with a lot of variability in the acoustic signal. For example, varying transmission channels, noise, speakers, and speaking styles are sources of such variabilities. Most of these variabilities are irrelevant for the speech recognition process and are actually sources of degraded recognition performance [Sankar & Lee 96]. From a more general perspective, this can be viewed as mismatch between training and testing condition of the ASR system. The training material of an SI-ASR system is usually chosen to contain a wide range of different acoustic conditions, for instance

- different speakers, speaking styles and accents,
- different transmission channels and room acoustics
- different types of microphones (close talking, far-field, headset)
- different levels and sources of ambient and channel noise

This collection of acoustic conditions in the training material is necessary to enable the ASR system to cope with the different conditions. On the other hand, this variety of acoustic conditions broadens the models trained from these training data. A given test utterance usually contains a specific combination of the acoustic conditions (i.e. one speaker with a specific vocal tract length and speaking style using a particular microphone with a particular noise level).

Examples of degraded performance caused by such mismatch are:

- speaker dependent systems significantly outperform speaker independent systems
- gender-dependent systems generally obtain better recognition results (on the corresponding gender) than gender-independent systems

- systems trained on data recorded over a mobile phone in a car perform significantly better on the same test environment than systems trained on data collected in an office using a microphone

A schematic view of these mismatches between training and test conditions is depicted in Fig. 4.1. The left side refers to the feature space of the acoustic vectors

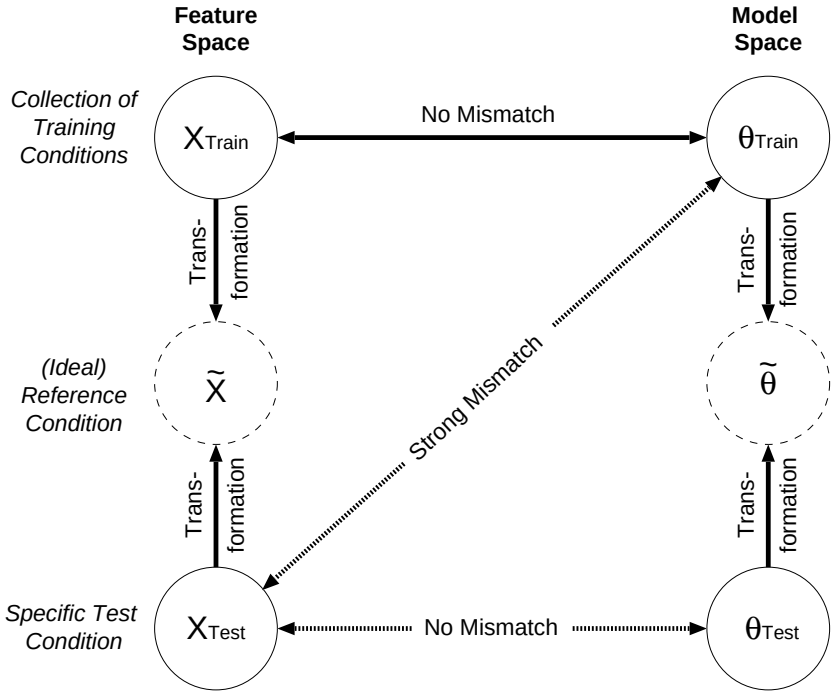

Figure 4.1: Schematic view of mismatch in training and test.

$x_1^T = x_1, \ldots, x_T$ and the right side to the model space of the corresponding acoustic models with model parameters θ. Three abstract data levels are distinguished:

- The training data ($\mathrm{X}_{\mathrm{Train}}$, first level), which contain a collection of different conditions. The model parameters θ_{Train} are estimated on the set $\mathrm{X}_{\mathrm{Train}}$ during training. θ_{Train} covers, to a certain degree, the conditions represented in the training data $\mathrm{X}_{\mathrm{Train}}$.

- A particular test utterance ($\mathrm{X}_{\mathrm{Test}}$, third level), which usually contains one specific condition only. In general, this exact condition has not been part of the training data, e.g. the specific speaker or background noise. If the acoustic model (θ_{Test}) had been trained on this specific test condition, there would have been no mismatch. This implies on the other hand a strong mismatch if feature vectors $\mathrm{X}_{\mathrm{Test}}$ are to be recognized with an acoustic model θ_{Train}.

- An (ideal) intermediate level ($\tilde{\mathrm{X}}$). At this level it is assumed that the variation in the acoustic signal caused by different conditions is removed by a transformation of the original feature vectors, e.g. Cepstral mean normalized feature vectors to eliminate different transmission channels, or vocal tract normalized feature vectors to eliminate the variations caused by different vocal tracts of the individual speakers. This level is an idealized view because the variations in the acoustic signal cannot be removed completely in practice.

Using this framework, speech recognition can be viewed as combination of feature vectors and acoustic models from those specific data levels. A mismatch is given if both spaces do not belong to the same level. For instance, in the case of non-adaptive acoustic modeling there is a strong mismatch between the test data $\mathrm{X}_{\mathrm{Test}}$ and the acoustic model θ_{Train}.

4.2 Adaptation and Normalization

Besides training condition-dependent acoustic models, several approaches have been developed to compensate for the mismatch described in the last Section. A schematic overview of adaptive acoustic modeling is depicted in Fig. 4.2. Again, this picture

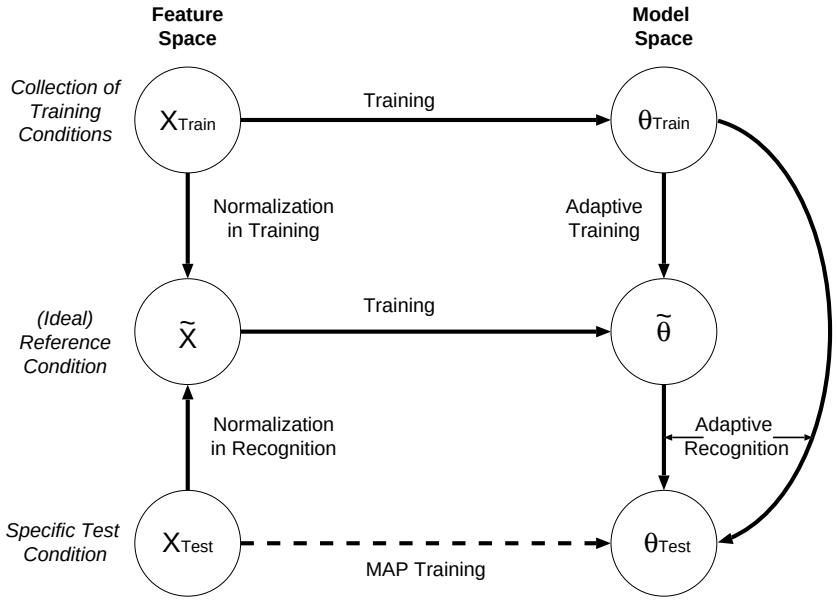

Figure 4.2: Overview of normalization (left side) and adaptation (right side) schemes.

shows the three abstract data levels as in Fig. 4.1 together with approaches to overcome the mismatch that is given for a combination of feature vectors X and

acoustic models θ from different levels. The mismatch can be reduced in the feature space (*normalization*) or in the model space (*adaptation*):

- In normalized acoustic modeling the variations in the acoustic signal caused by a specific source (e.g. different vocal tracts of individual speakers) are tried to be removed during signal analysis, yielding normalized acoustic vectors $\tilde{\mathrm{X}}$. As can be seen from Fig. 4.2, normalization approaches have to be applied to both the training ($\mathrm{X}_{\mathrm{Train}}$) *and* test data ($\mathrm{X}_{\mathrm{Test}}$) to gain maximum performance. If applied to only one of both, a slight mismatch remains (between $\mathrm{X}_{\mathrm{Test}}$ and $\tilde{\theta}$ if the training data are normalized only or between $\tilde{\mathrm{X}}$ and θ_{Train} if the test data are normalized only).

- Adaptation schemes modify the parameters of the acoustic model directly in order to reduce the mismatch. Provided that the applied adaptation scheme has sufficient degrees of freedom it is capable of reducing the mismatch between $\mathrm{X}_{\mathrm{Test}}$ and θ_{Train} by (ideally) transforming θ_{Train} into θ_{Test}. Thus adaptation schemes may be applied in recognition only and the additional benefit of adaptive training is small.

The current adaptation and normalization approaches can be categorized into two classes: the *maximum a-posteriori (MAP)* family and the *transformation* family.

4.2.1 MAP Family

Maximum a-posteriori adaptation (MAP) [Gauvain & Lee 94] follows the principle of Bayesian parameter estimation [Duda & Hart$^+$ 01]. In contrast to maximum-likelihood (ML) estimation where the parameter set θ is *fixed* but unknown, Bayesian estimation considers the parameters θ themselves as random variable drawn from a prior distribution $p(\theta|\tau)$ with so called hyper-parameters τ. Usually the hyper-parameters τ are estimated on training data $\{x\}$. In general, the dependency on θ has to be treated as hidden variable:

$$p\left(x_1^T|\tau(\{x\})\right) = \int d\theta \; p(x_1^T|\theta)\, p\big(\theta|\tau(\{x\})\big)$$

In MAP adaptation this integral is approximated by the maximum

$$p\left(x_1^T|\tau(\{x\})\right) = \max_{\theta} \; p(x_1^T|\theta)\, p\big(\theta|\tau(\{x\})\big)$$

and the adapted recognition is carried out with the MAP estimate

$$\theta_{\mathrm{map}} = \operatorname*{argmax}_{\theta} \; p(x_1^T|\theta)\, p\big(\theta|\tau(\{x\})\big). \tag{4.1}$$

used as parameters of the acoustic model:

$$[w_1^N]_{\mathrm{opt}} = \operatorname*{argmax}_{w_1^N} \left\{p(x_1^T|w_1^N;\theta_{\mathrm{map}}) \cdot p(w_1^N)\right\} \; .$$

For Gaussian mixture HMM the MAP estimate of the mean μ_i with prior $\mu_{i,0}$ is given as [Gauvain & Lee 94]

$$\mu_{i,\text{map}} = \frac{\lambda_i \mu_{i,0} + \sum_{t=1}^{T} \gamma_i(t) x_t}{\lambda_i + \sum_{t=1}^{T} \gamma_i(t)} \,, \tag{4.2}$$

where $\gamma_i(t)$ is the Gaussian occupation probability for density i at time t and λ_i is an adjustable parameter. In other words, the MAP estimate in Eq. (4.2) describes an interpolation of the prior mean $\mu_{i,0}$ and the empirical mean of the data x_1^T.

The advantage of MAP adaptation is that the parameters converge to those of a speaker-dependent system trained with ML, i.e. a MAP adapted system converges to a speaker-dependent system as the adaptation data increases. The main disadvantage is that MAP adaptation needs large amounts of adaptation data. As this is a local approach, only those parameters may be adapted which have been observed in the adaptation data. Up-to-date large vocabulary ASR systems make use of several 100.000 Gaussian densities, hence the number of unobserved parameters in the adaptation data will be huge for practical amounts of adaptation data. Several extensions to MAP have been proposed to overcome this disadvantage [Ahadi & Woodland 97, Shinoda & Lee 97].

4.2.2 Transformation Family

In contrast to directly adapting the parameters of the acoustic model, another approach is to apply a transformation to the original parameters. The advantage is that the transformation may depend on a few parameters only and that several (or even all) densities of the acoustic model may share the same transformation. The transformations are usually divided into two classes [Leggetter 95]:

- *Normalization* (left side in Fig. 4.2) tries to transform the feature vectors to a reference condition in which the effect causing the mismatch is (ideally) removed. An example of such a normalization scheme is vocal tract normalization (VTN) [Wakita 77, Eide & Gish 96, Lee & Rose 96, Wegmann & McAllaster$^+$ 96]. In a simplified model the human vocal tract is assumed as a uniform tube. A change in the length of the tube by a factor α^{-1} results in a scaling of the frequency axis by a factor α. Accordingly, the mismatch caused by different vocal tracts of the particular speakers is reduced by a scaling the frequency axis of the power spectrum (cf. Chapter 7). The basic idea is as follows: the frequency axis is scaled by a warping function g_a with a transformation parameter α

$$\begin{aligned} g_\alpha : [0,\pi] &\rightarrow [0,\pi] \\ \omega &\rightarrow \tilde{\omega} = g_\alpha(\omega) \,, \end{aligned} \tag{4.3}$$

where ω denotes the original frequency and $\tilde{\omega}$ the warped frequency. The warping function g_α is assumed to be invertible, i.e. strictly monotonic and continuous (see Fig. 4.3). The frequency $\omega = \pi$ corresponds to the Nyquist frequency and the domain and co-domain are chosen to conserve bandwidth and information contained in the original spectrum.

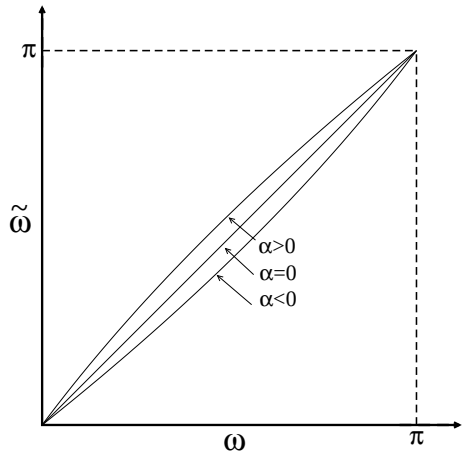

Figure 4.3: Example of a VTN warping function $\tilde{\omega} = g_\alpha(\omega)$ for different values of the warping parameter α.

All typical VTN approaches have in common that the warping function depends only on a few free parameters which control the amount of the frequency distortion. Even with only one free parameter (the warping factor α) to model the mismatch, VTN performs very well in a variety of recognition tasks. On the other hand, a few parameters allow to be estimated reliably on very little data which makes VTN a good choice for on-line speech recognition systems. More details on VTN warping are given in Chapter 6

- *Adaptation* (right side in Fig. 4.2) denotes the transformation of the parameters of the acoustic model, typically the parameters of the emission probability. An example of an adaptation approach is Maximum Likelihood Linear Regression (MLLR) [Leggetter & Woodland 95a] in which the mismatch is reduced by an affine transformation of the mean vectors of the emission probability distribution of the HMM model:

$$\hat{\mu}_{s,r} = \boldsymbol{A}_{s,r}\,\mu_s + b_{s,r} \tag{4.4}$$

where s denotes the HMM state and r the speaker. The matrix $\boldsymbol{A}_{s,r}$ is estimated via maximum likelihood, given adaptation data (x_1^T, w_1^N) or using a preliminary transcription $\tilde{w}_1^N$ of a first recognition pass (which usually will contain recognition errors).

An overview of current adaptation and normalization schemes is given in [Woodland 01]. Adaptation approaches work very well if applied in recognition only, i.e. without modifying the training part of the ASR system as they change the parameters θ_{train} trained on the collection of training conditions straightly to the parameters θ_{test} that are suitable for the specific test condition. Minor additional improvements can be obtained if adaptation schemes are also used in training (speaker adaptive training SAT) [Anastasakos & McDonough$^+$ 96]. In contrast, normalization schemes give best results if applied both in training and test. If used in recognition only, a minor mismatch between the normalized feature vectors $\tilde{x}_{\text{test}}$ and θ_{train} remains, which lowers the recognition performance. Applying normalization or adaptation only in training is disadvantageous as the resulting acoustic models do not contain enough flexibility to cope with unseen conditions [Welling & Kanthak$^+$ 99, Gales 01].

4.2.3 Text Dependency

Besides the distinction of the space the transformation is applied to, the different approaches may be classified according to their text or class dependency. Some approaches need a (word level) transcription or class labeling of the adaptation data (i.e. the data where the parameters are estimated on). Examples of such *text dependent* approaches are MLLR and MAP. VTN may be applied text-dependent like the common 2-pass approach as described in [Lee & Rose 96], where a word-level transcription is needed to estimate the parameter α, or text-independent, where α is estimated using text-independent Gaussian mixture models (fast-VTN, [Wegmann & McAllaster$^+$ 96, Welling & Haeb-Umbach$^+$ 98]). Typical text-independent approaches to reduce the mismatch are Cepstral mean and variance normalization, stochastic feature space matching [Sankar & Lee 95] or histogram normalization [Molau & Keysers$^+$ 02].

4.3 Mathematical Framework for Adaptive Acoustic Modeling

Adaptation and normalization are commonly viewed as different techniques to reduce the mismatch between training and testing conditions, e.g. in [Woodland 01], and have so far been treated separately. In fact, the terms adaptation and normalization are not orthogonal and the approaches can be described together in the same mathematical framework. In this Section a common mathematical framework will be developed and it will been shown that in terms of Bayes' decision rule there is no difference between these two approaches. However, the terms normalization and adaptation will be used throughout this work as in practice there maybe still some differences.

According to Eq. (1.2) the acoustic model distribution $p(x_1^T|w_1^N)$ in an HMM framework is given as

$$p(x_1^T|w_1^N;\theta) = \sum_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t;w_1^N;\theta) \cdot p(s_t|s_{t-1};w_1^N) \ ,$$

where θ denotes the (set of) parameters of the emission probability distribution, e.g. mean vector and covariance matrix for Gaussian distributions. To model the mismatch between different acoustic conditions described in the last Section a new condition-dependent parameter set α is introduced[1]:

$$p(x_1^T|w_1^N;\theta) \rightarrow p(x_1^T|w_1^N;\theta,\alpha) \ . \tag{4.5}$$

Typical examples of such condition-dependent parameter sets are the warping factor used in VTN (cf. Section 5.2.4) as a single parameter or the elements of the transformation matrix used in MLLR adaptation (cf. Section 5.2.1). For simplicity only a single parameter will be considered in the following, the extension to a set of parameters is obvious. In the HMM framework, the adaptation is typically applied to the parameters of the emission probability distribution only:

$$p(x_1^T|w_1^N;\theta,\alpha) = \sum_{s_1^T:w_1^N} \prod_{t=1}^{T} p(x_t|s_t;w_1^N;\theta,\alpha) \cdot p(s_t|s_{t-1};w_1^N) \ . \tag{4.6}$$

As the condition-dependent parameter is usually unobserved, it is considered as hidden variable:

$$p(x_1^T|w_1^N;\theta) = \int d\alpha \ p(x_1^T,\alpha|w_1^N;\theta) \tag{4.7}$$

$$= \int d\alpha \ p(\alpha|w_1^N;\theta) \cdot p(x_1^T|w_1^N;\theta,\alpha) \ . \tag{4.8}$$

Thus, Bayes' decision rule with adaptive acoustic modeling using the language model $p(w_1^N)$ becomes

$$[w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}} \left\{ p(w_1^N) \cdot \int d\alpha \ p(\alpha|w_1^N;\theta) \cdot p(x_1^T|w_1^N;\theta,\alpha) \right\} \ .$$

Often the integral is approximated by the maximum

$$p(x_1^T|w_1^N;\theta) \cong \max_{\alpha} \left\{ p(\alpha|w_1^N;\theta) \cdot p(x_1^T|w_1^N;\theta,\alpha) \right\}$$

[1] the same symbol p are for simplicity, although the functional form may change as well

and in this case Bayes' decision rule reads

$$[w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}} \left\{ p(w_1^N) \cdot \max_{\alpha} \left\{ p(\alpha|w_1^N;\theta) \cdot p(x_1^T|w_1^N;\theta,\alpha) \right\} \right\} \ . \tag{4.9}$$

The prior distribution $p(\alpha|w_1^N;\theta)$ is often assumed to be non-informative (i.e. uniform) and is hence neglected.

To obtain a model for practical applications, the exact dependence of the model on the parameter α must be specified. The aim of adaptive acoustic modeling is to move from model $p(x_1^T|w_1^N;\theta)$ with mismatch to a model $p(x_1^T|w_1^N;\theta,\alpha)$ with reduced (ideally removed) mismatch by means of the mismatch parameter α. In virtually all approaches the functional form of the probability density function is kept fixed and the dependence on the parameter α is realized via transformations[2]. These transformations can be applied in two ways:

- Transformation of the observations x_1^T (normalization):

$$\begin{aligned} x_1^T \rightarrow {x_\alpha}_1^T &= f_\alpha(x_1^T) \qquad \text{(element wise)} \\ p(x_1^T|w_1^N;\theta,\alpha) &= p(f_\alpha(x_1^T)|w_1^N;\theta) \cdot \left| \frac{df_\alpha(x_1^T)}{dx} \right| \end{aligned} \tag{4.10}$$

  where the last term is the Jacobian determinant of the transformation. A prototypical example of a transformation of the observation vector is VTN. The Jacobian determinant can be omitted if no direct comparison of probability values is carried out with differently "normalized" distributions. In a typical VTN approach the warping factor α is chosen by comparing the scores obtained by a forced alignment of the same utterance warped with a set of discrete warping factors. Hence, the Jacobian determinant needs to be taken into account [Sankar & Lee 96].
  Often the Jacobian determinant is assumed to be flat as function of α. Thus it is approximated to be independent of α and neglected. In virtually all experimental studies of VTN the Jacobian determinant has been neglected since it can hardly be calculated in the usual VTN technique with explicitly warping the frequency axis of the spectrum during signal analysis. Whether this approximation is justified has not been analyzed in detail so far. In [McDonough 00] it was shown that for the case of all-pass transforms the Jacobian determinant is important. With the approach presented in this work it is now possible to calculate the Jacobian determinant for VTN (see Chapter 8) and to investigate the importance of the Jacobian determinant for general warping functions.

[2] in this sense even MAP may be viewed as transformation of the parameters of the prior distribution, cf. Eq. (4.2)

- Transformation of the model parameters θ (adaptation):
  adapting the model parameters can often be formulated as an inverse transformation applied to the acoustic vectors (cf. Chapter 5). To simplify the notation, the same functional form f_α is used for both transformations:

  $$\theta \rightarrow \theta_\alpha = f_\alpha^{(-1)}(\theta)$$

  for which the probability distribution becomes

  $$p(x_1^T|w_1^N;\theta,\alpha) = p(x_1^T|w_1^N;f_\alpha^{(-1)}(\theta)) \; . \tag{4.11}$$

  In this case the random variable x_1^T is not altered and thus no Jacobian determinant is required. A prototypical example of this transformation is a linear transformation of the model parameters with a matrix $\mathbf{A}(\alpha)$ (MLLR):

  $$f_\alpha : \quad \theta \rightarrow \theta_\alpha = \mathbf{A}(\alpha) \cdot \theta$$

In summary, adaptation and normalization can be described in the same mathematical framework. A schematic overview is depicted in Fig. 4.4. Although adapta-

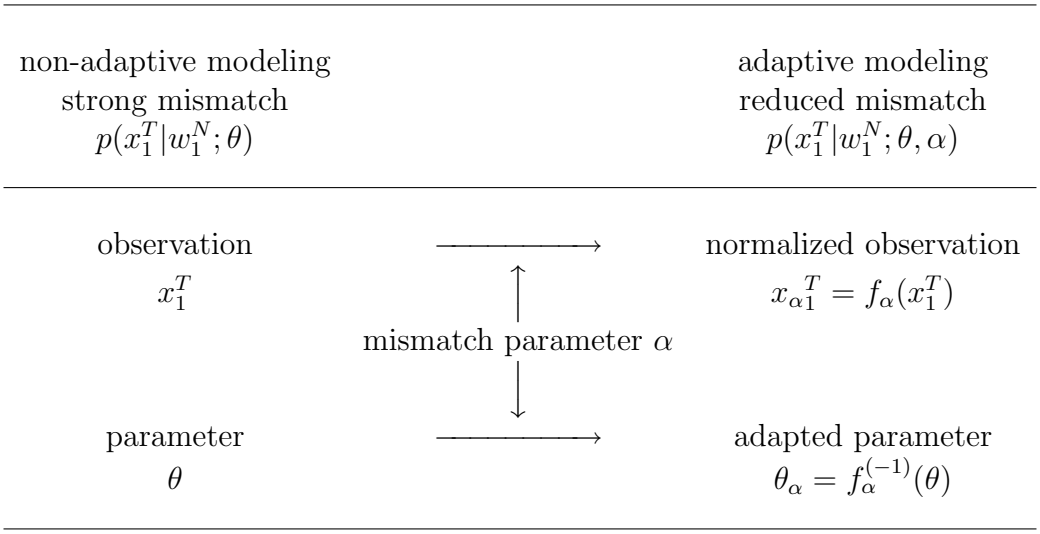

Figure 4.4: Schematic overview of adaptive acoustic modeling. Note simplification: the same functional form $f_\alpha(\cdot)$ has been used for both transforming the observations x_1^T and parameters θ.

tion and normalization are equivalent according to this framework, both presented ways of dealing with mismatch are relevant in practice. First it is necessary to find a suitable transformation function f_α for a given mismatched condition. Often it is

easier to formulate an appropriate transformation function either in the feature or in the model space. In some cases, the corresponding inverse transformation function may be hardly formulated. In practical applications it is relevant to estimate the condition dependent (sets of) parameters α reliably on limited amounts of data, which often is easier in one space.
It will be shown later that there is a strong interdependence of two widely used normalization (VTN) and adaptation (MLLR) techniques if Gaussian emission probabilities are used (cf. Section 7.7).

4.4 Summary

In this Chapter normalization and adaptation approaches, which have so far been treated separately, have been presented for the first time in a unified view. It has been shown that, in terms of Bayes' decision rule, both adaptation and normalization are interchangeable. Based on that, a consistent common mathematical framework for adaptive acoustic modeling could be achieved.

Additionally, the most important adaptation/normalization families together with typical representatives have been briefly introduced.

Chapter 5

Unified View of Linear Transformations and Vocal Tract Normalization

5.1 Motivation

The application of linear transformations to either the acoustic feature vector or the parameters of the acoustic model has a long tradition is speech recognition. For instance, an approach to use a linear transformation of the acoustic feature vectors to enhance the separability of syllables has already been proposed in [Hunt 79]. The approaches described in [Jaschul 82] and [Class & Kaltenmeier$^+$ 90] led to the nowadays widely used MLLR speaker adaptation [Leggetter & Woodland 95a] (cf. Section 9).

A confusing abundance of linear transformations has been proposed by now. Examples of speaker independent transformations are linear discriminant analysis (LDA) [Brown 87, Doddington 89], heteroscedastic discriminant analysis (HDA) [Kumar & Andreou 98], semi-tied covariances (STC) [Gales 99], maximum likelihood linear transformations (MLLT) [Gopinath 98] or extended MLLT (EMLLT) [Olsen & Gopinath 02]. Another group of transformations contains the speaker dependent transformations MLLR [Leggetter & Woodland 95a], variance adaptation [Gales & Woodland 96], "constrained" MLLR (C-MLLR), feature space MLLR (F-MLLR) [Gales 98] or vocal tract normalization [Eide & Gish 96, Lee & Rose 96], which will be shown in Chapter 7 to be a linear transformation as well.

While the individual approaches have been studied in great detail in literature, the relationship between them has been neglected and never been discussed thoroughly. As a matter of fact, there exist close connections between the approaches listed above, both within and between the two groups.

This Chapter will start with a review of the most important linear transformations in speech recognition and it will be shown that all these linear transformations can be formally described in a common framework. A unified view of linear transformations will be presented. Based on that common framework, the manifold

links between the approaches will be elaborated and some approaches will turn out to be a special case of another or are even mathematically equivalent.

Hidden Markov Models (HMM) have been established as de-facto standard for state-of-the-art speech recognition systems [Baker 75, Rabiner 89]. In most systems the emission probabilities are modeled by mixtures of Gaussian probability distributions

$$p(x|s) = \sum_{l=1}^{L_s} c_{sl}\, p(x|\mu_{sl}, \mathbf{\Sigma}_s) \;, \tag{5.1}$$

where x denotes the observation vector, s the HMM state and c_{sl} the weight of the mixture component l. A transformation of the mixture weights is seldomly used in speech recognition systems, so the following discussion will focus on the transformation of the mean and/or covariances of the single Gaussian probability distributions:

$$\hat{\mu} = \boldsymbol{A}\,\mu + b \qquad \hat{\mathbf{\Sigma}} = \boldsymbol{H}\,\mathbf{\Sigma}\,\boldsymbol{H}^\top \tag{5.2}$$

where $\boldsymbol{A}$ and $\boldsymbol{H}$ are $n \times n$ transformation matrices and n is the dimension of the acoustic feature vector. Thus, the transformed Gaussian emission probability for the HMM state s becomes

$$p(x|\mu_s, \mathbf{\Sigma}_s; \mathbf{A}, b, \boldsymbol{H}) = \frac{1}{\sqrt{|\det(2\pi \boldsymbol{H}\,\mathbf{\Sigma}_s\,\boldsymbol{H}^\top)|}} \exp\left(-\frac{1}{2}(x - \boldsymbol{A}\,\mu_s - b)^\top \left(\boldsymbol{H}\,\mathbf{\Sigma_s}\,\boldsymbol{H}^\top\right)^{-1} (x - \boldsymbol{A}\,\mu_s - b)\right) \tag{5.3}$$

In the following two Sections the linear transformations will be introduced and motivated as presented by the original authors. A unifying view and the interconnection between them will be discussed thereafter.

The linear transformations can be divided into speaker dependent and speaker independent transformations. The main difference is the data on which the transformations are estimated.

5.2 Speaker Dependent Transformations

Speaker dependent transformations are mostly used to adapt a speaker-independent system to a new speaker. The transformations are estimated on a certain amount of data from the new speaker and the so obtained transformations are applied to the speaker-independent model parameters, yielding a speaker-dependent, or better speaker-*adapted* system. The adaptation data may be collected beforehand or the

data to be recognized itself may be used in a two-pass recognition approach or by iterative adaptation.

The most flexible transformation approach would estimate an individual transformation for each HMM state. Due to usually limited amount of data to estimate the parameters on, regression classes $c = 1, \ldots, C$ are defined for which the transformations are estimated [Gales 96]. Thus several HMM states share the same transformation. The granularity of these regression classes is dependent on the actual amount of adaptation data. In order to simplify the equations, the presentation starts with one transformation for each state and regression classes will be considered later.

Using a maximum likelihood parameter estimation, the speaker dependent transformations are obtained by maximizing Eq. (5.3) for each speaker $r = 1, \ldots, R$:

$$(\boldsymbol{A}_{s,r}^{\text{opt}}, b_{s,r}^{\text{opt}}, \boldsymbol{H}_{s,r}^{\text{opt}}) = \underset{(\boldsymbol{A}_{s,r}, b_{s,r}, \boldsymbol{H}_{s,r})}{\operatorname{argmax}} \left\{ \sum_{t=1}^{T_r} \sum_{s'=1}^{S} \gamma_{s'}(t) \left[\log \left(\frac{1}{|\det \boldsymbol{H}_{s',r}\, \boldsymbol{\Sigma}_{s'}\, \boldsymbol{H}_{s',r}^{\top}|} \right) \right. \right.$$
$$\left. \left. - \left((x_t - \boldsymbol{A}_{s',r}\, \mu_{s'} - b_{s',r})^{\top} \left(\boldsymbol{H}_{s',r}\, \boldsymbol{\Sigma}_{s'}\, \boldsymbol{H}_{s',r}^{\top} \right)^{-1} (x_t - \boldsymbol{A}_{s',r}\, \mu_{s'} - b_{s',r}) \right) \right] \right\} \quad (5.4)$$

Some specific realizations have been studied in literature and will be discussed in the following.

5.2.1 Maximum Likelihood Linear Regression

Maximum likelihood linear regression (MLLR) [Leggetter & Woodland 95a] denotes an affine transformation of the means of the emission probability, estimated by maximum likelihood:

$$\hat{\mu}_{s,r} = \boldsymbol{A}_{s,r}\, \mu_s + b_{s,r} \; . \quad (5.5)$$

The covariances are left unchanged, i.e. the transformation matrix $\boldsymbol{H}$ is set to $\boldsymbol{H} = \boldsymbol{1}$. For notational convenience the affine transformation of Eq. (5.5) is rewritten in the form

$$\hat{\mu}_{s,r} = \boldsymbol{W}_{s,r}\, \xi_s \quad (5.6)$$

where ξ_s denotes the extended mean vector

$$\xi_s = [1\ \mu_s^{\top}]^{\top} \quad (5.7)$$

and $\boldsymbol{W}_{s,r}$ is the $n \times (n+1)$-matrix $[b_{s,r}\, \boldsymbol{A}_{s,r}]$. The adaptation matrix $\boldsymbol{W}_{s,r}$ is estimated by maximum likelihood, given adaptation data (x_{r1}^{T}, w_{r1}^{N}) consisting of the sequence of observation vectors x_{r1}^{T} and the word sequence w_{r1}^{N} for speaker r:

$$\boldsymbol{W}_{s,r}^{\text{ML}} = \underset{\boldsymbol{W}_{s,r}}{\operatorname{argmax}}\, p(x_{r1}^{T} | \mu_s, \boldsymbol{\Sigma}_s, \boldsymbol{W}_{s,r}; w_{r1}^{N}) \quad (5.8)$$

The word sequence ${w_r}_1^N$ may be either the true spoken sequence (supervised adaptation) or been obtained by a previous recognition pass (unsupervised adaptation). The latter will usually contain recognition errors and thus supervised adaptation is usually superior to unsupervised adaptation.

The maximization of Eq. (5.8) is carried out using the expectation-maximization (EM) algorithm [Dempster & Laird$^+$ 77] by maximizing the auxiliary function

$$Q(\lambda, \bar{\lambda}) = \sum_{s_1^T : {w_r}_1^N} p({x_r}_1^T | {w_r}_1^N, \lambda) \log p({x_r}_1^T | {w_r}_1^N, \bar{\lambda}) \tag{5.9}$$

where λ denotes the original set of model parameters and $\bar{\lambda}$ the updated (i.e. adapted) model set of the iterative optimization scheme. Omitting terms which are independent of $\boldsymbol{W}$, the following equation needs to be solved:

$$\boldsymbol{W}_{s,r}^{\mathrm{ML}} = \underset{\boldsymbol{W}_{s,r}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T} \gamma_s(t) (x_{rt} - \boldsymbol{W}_{s,r}\, \xi_s)^\top \boldsymbol{\Sigma}_s^{-1} (x_{rt} - \boldsymbol{W}_{s,r}\, \xi_s) \right\} . \tag{5.10}$$

When using regression classes, an additional summation must be performed over all HMM states belonging to the same regression class c. For diagonal covariance models, which are used in virtually all speech recognition systems due to computational efficiency and data sparseness problems, a closed-form solution of Eq. (5.4) can be obtained. Taking the derivative w.r.t. $\boldsymbol{W}_{c,r}$ and equating to zero yields a row-wise solution for $\boldsymbol{W}_{c,r}^{\mathrm{ML}}$ [Leggetter & Woodland 95a]:

$$W_{c,r}^{(i)} = Z_c^{(i)} {\boldsymbol{G}_c^{(i)}}^{-1} \tag{5.11}$$

with

$$\boldsymbol{G}_c^{(i)} = \sum_{m=1}^{M} \frac{1}{\sigma_{s_m i}^2} \xi_{s_m} \xi_{s_m}^\top \sum_{t=1}^{T} \gamma_{s_m}(t) \tag{5.12}$$

$$Z_c^{(i)} = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{s_m}(t) \frac{1}{\sigma_{s_m i}^2} x_{rt} \xi_{s_m}^\top \tag{5.13}$$

where $W_{c,r}^{(i)}$ and $Z_c^{(i)}$ denote the i-th row-vector of $\boldsymbol{W}_{c,r}^{\mathrm{ML}}$ and $\boldsymbol{Z}_c$, respectively, and $\sigma_{s\,i}^2$ is the i-th diagonal element of $\boldsymbol{\Sigma}_s$. The sum over m runs over all states $\mathcal{S}_c = \{s_1, \ldots, s_m, \ldots, s_M\}$ which belong to class c. A solution for the full covariance case is given in [Gales & Woodland 96].

MLLR provides a powerful and easy to implement adaptation technique which is used in most of today's state-of-the-art speech recognition systems, e.g. [Saon & Zweig$^+$ 03, Nguyen & Rigazio$^+$ 03, Evermann & Chan$^+$ 04, Schwartz & Colthurst$^+$ 04]. Typical improvements of MLLR adaptation range between 10% and 20% relative reduction in word error rate.

5.2.2 Variance Adaptation

Variance adaptation [Gales 98] denotes a speaker dependent transformation of the covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{s,r} = \boldsymbol{H}_{c,r}\, \boldsymbol{\Sigma}_s\, \boldsymbol{H}_{c,r}^{\top}\,. \tag{5.14}$$

Generally, a variance adaptation can be applied independently from a transformation of the means. In practice, however, both approaches have been always used in combination. It has been found that transforming the means is more important than transforming the covariances, which gives only a small additional improvement [Gales & Woodland 96].

5.2.3 Constrained MLLR

A special combination of mean and variance transformation is *constrained* MLLR (C-MLLR), where the transformation matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ are forced to be equal, i.e. mean and variances are transformed using the same transformation matrix [Gales 98]. The advantage of this approach is that the transformation is equivalent to a transformation of the feature vector instead of the model parameters and thus allows for an efficient implementation:

$$\begin{aligned} & p(x|\mu_s, \boldsymbol{\Sigma}_s; \boldsymbol{A}) = \\ & \frac{1}{\sqrt{|\det(2\pi \boldsymbol{A}\, \boldsymbol{\Sigma}_s\, \boldsymbol{A}^{\top})|}} \exp\left(-\frac{1}{2}(x - \boldsymbol{A}\,\mu_s - b)^{\top} \left(\boldsymbol{A}\, \boldsymbol{\Sigma_s}\, \boldsymbol{A}^{\top}\right)^{-1} (x - \boldsymbol{A}\,\mu_s - b)\right) \\ = & \frac{|\det \boldsymbol{A}|}{\sqrt{|\det(2\pi \boldsymbol{\Sigma}_s)|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{A}^{-1}(x-b) - \mu_s\right)^{\top} \boldsymbol{\Sigma}_s^{-1} \left(\boldsymbol{A}^{-1}(x-b) - \mu_s\right)\right) \end{aligned} \tag{5.15}$$

Therefore C-MLLR is sometimes called *feature space* MLLR (F-MLLR). Experimental studies have revealed little differences between mean and variance adaptation without constraint, i.e. using different matrices for mean and variance, and C-MLLR. As C-MLLR can be implemented as simple transformation of the observation vector this approach has prevailed, especially for speaker adaptive training, c.f. [Saon & Zweig^{+} 03, Evermann & Chan^{+} 04, Schwartz & Colthurst^{+} 04].

5.2.4 Vocal Tract Normalization

Vocal tract normalization (VTN) is discussed in more detail in Section 4.2.2 and Chapter 7. VTN has been originally formulated as a warping of the frequency axis of the power spectrum to compensate for the different vocal tract lengths of the individual speakers. It will be shown in Chapter 7 that vocal tract normalization (VTN) can be expressed as linear transformation of the Cepstral vector. The

transformation matrix is dependent on a single parameter only, the warping factor α. Thus, VTN can be viewed as a highly restricted C-MLLR transformation, i.e. with a C-MLLR matrix $\boldsymbol{A}_r(\alpha)$ with only one free parameter α. This result, which will be derived analytically in Chapter 7, has been experimentally observed in [Uebel & Woodland 99], where the authors found that improvements obtained from VTN and C-MLLR were not additive. Thus, VTN is a special realization of C-MLLR.

5.2.5 Extensions to MLLR

Several extensions to the MLLR approach have been presented in literature. In [Chesta & Siohan$^+$ 99] and [Chou 99] the estimation of the transformation matrices is based on a maximum a-posteriori (MAP) criterion by the use of a prior distribution for the transformation matrices (MAPLR). In [Gunawardana & Byrne 01] a discounted likelihood optimization criterion has been proposed for estimating the transformation matrix (DLLR). Both MAPLR and DLLR improve the robustness of the adaptation when only small amounts of adaptation data are available. Recently, also discriminative techniques have been used as optimization criterion, e.g. [Wallhoff & Willet$^+$ 00, Gao & Ramabhadran$^+$ 00, Uebel & Woodland 01, Doumpiotis & Tsakalidis$^+$ 04].

As this work focuses an the general attributes of linear transformations, those extensions will not be discussed in detail.

5.2.6 Summary

The speaker dependent transformations can be classified according to the following scheme:

Table 5.1: Overview of speaker dependent linear transformations based on Eq. (5.3). The subscript r denotes a speaker dependency and the subscript c a dependency on regression (adaptation) classes.

transformation	$\boldsymbol{A}$	$\boldsymbol{H}$	remark
MLLR	$\boldsymbol{A}_{c,r}$	$\mathbf{1}$	
variance adaptation	$\mathbf{1}$	$\boldsymbol{H}_{c,r}$	
mean & var. adapt.	$\boldsymbol{A}_{c,r}$	$\boldsymbol{H}_{c,r}$	
C-MLLR/F-MLLR	$\boldsymbol{A}_r$	$\boldsymbol{A}_r$	
VTN	$\boldsymbol{A}_r(\alpha)$	$\boldsymbol{A}_r(\alpha)$	C-MLLR with restricted matrix (only one free parameter α)

Usually C-MLLR and VTN are applied using one global transformation whereas

for MLLR and variance adaptation regression classes are defined for which different transformation matrices are estimated.

5.3 Speaker Independent Transformations

Speaker independent linear transformations are typically employed to transform the feature space into a (sub)space where certain model assumptions are more appropriate than in the original space. For instance, the well-known LDA transformation [Fisher 36, Duda & Hart$^+$ 01] finds a subspace in which class separability is enhanced for globally pooled diagonal covariance modeling and arranges the components of the feature vector in the new space according to their importance for classification.

Two different classes of linear speaker independent transformations are used in speech recognition systems: One class contains a dimensionality reduction like LDA [Brown 87, Doddington 89] and HDA [Kumar & Andreou 98]. Those transformations aim at projecting the feature space into one of lower dimension by selecting the most important directions in the feature space. The other class of transformations like STC, MLLT or EMLLT maintains the dimension of the feature space. The aim of STC for instance is to find a feature space in which the assumption of diagonal covariance matrices is more appropriate than in the original space. In the following the differences and similarities between the approaches will be discussed in detail.

Using a maximum likelihood parameter estimation, the estimation equation is very similar to that for speaker dependent transformations (Eq. (5.4), presented in Section 5.2). The only difference is that the transformations are obtained by maximizing Eq. (5.3) jointly for all data instead of using data from each speaker separately:

$$(\boldsymbol{A}_s^{\text{opt}}, b_s^{\text{opt}}, \boldsymbol{H}_s^{\text{opt}}) = \underset{(\boldsymbol{A}_s, b_s, \boldsymbol{H}_s)}{\operatorname{argmax}} \left\{ \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{s'=1}^{S} \gamma_{s'}(t) \left[\log \left(\frac{1}{|\det \boldsymbol{H}_{s'} \boldsymbol{\Sigma}_{s'} \boldsymbol{H}_{s'}^{\top}|} \right) - \left((x_t - \boldsymbol{A}_{s'} \mu_{s'} - b_{s'})^{\top} \left(\boldsymbol{H}_{s'} \boldsymbol{\Sigma}_{s'} \boldsymbol{H}_{s'}^{\top} \right)^{-1} (x_t - \boldsymbol{A}_{s'} \mu_{s'} - b_{s'}) \right) \right] \right\} \quad (5.16)$$

5.3.1 Semi-tied Covariances and MLLT

The idea of semi-tied covariances (STC) [Gales 97, Gales 99] is to find a global transformation $\boldsymbol{H}$ such that the resulting covariance matrices $\boldsymbol{\Sigma}_s$ of the emission probabilities are as close to diagonal as possible. The same approach has also been published in [Gopinath 98] and is referred to as Maximum Likelihood Linear Transformation (MLLT).

As known from linear algebra, it is always possible to simultaneously diagonalize two matrices if each matrix can be diagonalized individually (generalized Eigenvalue problem). With more than two matrices (as usual in today's speech recognition systems with about several hundred thousand Gaussian distributions), a simultaneous diagonalization can be obtained only approximatively. The goal is to find a transformation such that the error of using diagonal covariance models is minimized. Thus, the resulting $(D \times D)$ covariance matrix is given as

$$\hat{\boldsymbol{\Sigma}}_s = \boldsymbol{H}\, \boldsymbol{\Sigma}_s^{\text{diag}}\, \boldsymbol{H}^\top = \sum_{d=1}^{D} \sigma_{sd}^2 h_d h_d^\top \tag{5.17}$$

where $\boldsymbol{\Sigma}_s^{\text{diag}}$ is modeled as diagonal matrix of size $(D \times D)$, D is the dimension of the acoustic feature vector, σ_{sd}^2 is the d-th entry of $\boldsymbol{\Sigma}_s^{\text{diag}}$ and h_d is the d-th row vector of the $(D \times D)$-matrix $\boldsymbol{H}$, which is not necessarily orthogonal. In other words, the state-specific covariance matrix $\boldsymbol{\Sigma}_s$ is obtained by multiplying the state-specific *diagonal* covariance matrix $\boldsymbol{\Sigma}_s^{\text{diag}}$ with the state-*independent* transformation matrix $\boldsymbol{H}$.

Semi-tied covariance modeling has the advantage to overcome the limitation of diagonal covariance modeling without the enormous increase in the number of parameters that comes with full covariance modeling.

The transformation $\boldsymbol{H}^{\text{opt}}$ as well as the means μ_s^{opt} and covariances $\boldsymbol{\Sigma}_{\boldsymbol{s}}^{\text{diag}^{\text{opt}}}$ are obtained by maximum likelihood estimation. The optimization equation (5.16) reads for STC modeling

$$\begin{aligned}(\mu_s^{\text{opt}}, \boldsymbol{\Sigma}_{\boldsymbol{s}}^{\text{diag}^{\text{opt}}}, \boldsymbol{H}^{\text{opt}}) = \underset{(\mu_s, \boldsymbol{\Sigma}_s^{\text{diag}}, \boldsymbol{H})}{\text{argmax}} \Bigg\{ \sum_{t=1}^{T} \sum_{s'=1}^{S} \gamma_{s'}(t) \Bigg[\log \left(\frac{1}{|\det \boldsymbol{H}\, \boldsymbol{\Sigma}_{s'}^{\text{diag}}\, \boldsymbol{H}^\top|} \right) \\ - \left((x_t - \mu_{s'})^\top \left(\boldsymbol{H}\, \boldsymbol{\Sigma}_{s'}^{\text{diag}}\, \boldsymbol{H}^\top \right)^{-1} (x_t - \mu_{s'}) \right) \Bigg] \Bigg\} \quad . \end{aligned} \tag{5.18}$$

The resulting estimation equations can be found in [Gales 99]. STC modeling is closely related to variance adaptation presented in the previous Section. Comparing Eq. (5.4) and (5.18) shows that the estimation equations for variance adaptation and STC are nearly identical. The main difference are the data which are used for the optimization. While the STC transformation matrix is typically estimated from the training data, the transformation matrices for variance adaptation are estimated from speaker specific adaptation data (which is often identical to the test data for unsupervised adaptation).

5.3.2 EMLLT

Extended Maximum Likelihood Linear Transformation (EMLLT) [Olsen & Gopinath 02] is an extension to STC/MLLT. For a STC model, the

covariance matrix for each state $\hat{\mathbf{\Sigma}}_s$ has only D independent components (c.f. Eq. (5.17)):

$$\hat{\mathbf{\Sigma}}_s^{\text{STC}} = \sum_{d=1}^{D} \sigma_{sd}^2 h_d h_d^\top \quad . \tag{5.17}$$

Eq. (5.17) describes the composition of the covariance matrix $\hat{\mathbf{\Sigma}}_s^{\text{STC}}$ as a sum of D (D denotes the dimension of the acoustic feature vector) rank-one matrices $h_d\, h_d^\top$, i.e. $\hat{\mathbf{\Sigma}}_s^{\text{STC}}$ is generated by the basis $\{h_d h_d^\top\}_{d=1}^{D}$.

The idea of EMLLT is to increase the *number* of rank-one matrices $h_d h_d^\top$, which build the basis, while keeping the dimension of each vector h_d fixed to D:

$$\hat{\mathbf{\Sigma}}_s^{\text{EMLLT}} = \sum_{d=1}^{\Delta} \sigma_{sd}^2 h_d h_d^\top \quad \text{with } D \leq \Delta \leq D(D+1)/2 \; . \tag{5.19}$$

In other words, $\hat{\mathbf{\Sigma}}_s$ now consists of Δ independent components (instead of D for STC modeling) and the transformation matrix $\boldsymbol{H}$ is now of dimension $D \times \Delta$. The main differences of STC and EMLLT covariance modeling are schematically depicted in Fig. 5.1.

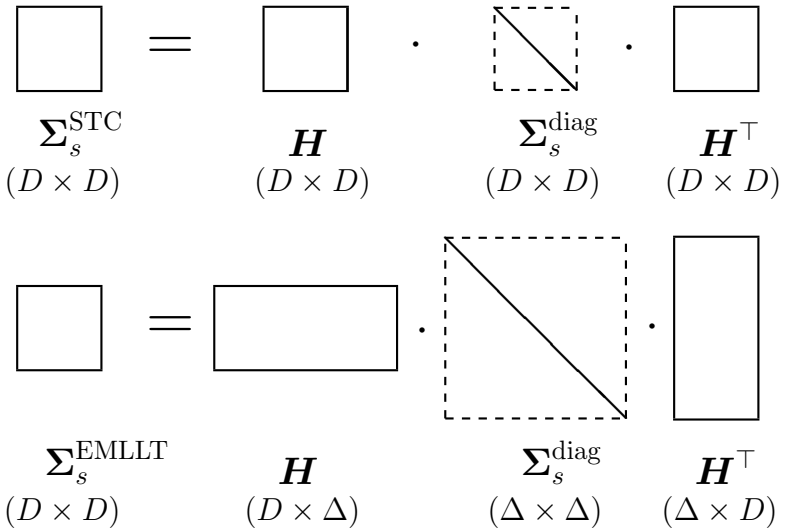

Figure 5.1: Schematic differences of STC and EMLLT modeling. The terms in parentheses denote the dimensions of the matrices.

EMLLT is equal to a full covariance modeling for $\Delta = D(D+1)/2$ if the vectors h_d are chosen in the following way:

$$h_d = \begin{cases} e_d & \text{for } d \leq D \\ e_i + e_j & \text{for } D < d \leq \Delta, \text{ with } i, j \in [1, D] \text{ and } i < j \end{cases} \tag{5.20}$$

e_k denotes the unit vector in direction k. The ML estimate of $\boldsymbol{H}$ is again given by Eq. (5.16) or Eq. (5.18), respectively. It should be noted that $\boldsymbol{H}$ is not invertible in this formulation and a rewriting similar to Eq. (5.15) is not possible. After introducing the HDA approach, it will be shown that there exists a close relationship between STC, EMLLT and HDA.

5.3.3 HDA and LDA

Heteroscedastic Discriminant Analysis (HDA) [Kumar & Andreou 98] and Linear Discriminant Analysis (LDA) [Fisher 36, Duda & Hart$^+$ 01] are examples of dimensionality reducing transformations. Dimensionality reducing transformations project a feature space of dimension $\mathcal{D}$ into a lower dimensional one with dimension $D < \mathcal{D}$. The objective is to reduce the dimensionality of the feature space to care for data sparseness problems that may occur in high dimensional feature spaces while retaining a maximum of class separability information:

$$\begin{aligned} F_{\boldsymbol{\Theta}_D} : \mathbb{R}^{\mathcal{D}} &\rightarrow \mathbb{R}^D \quad (D < \mathcal{D}) \\ \tilde{x} &\rightarrow x = \boldsymbol{\Theta}_D \tilde{x} \ , \end{aligned} \tag{5.21}$$

where $\boldsymbol{\Theta}_D$ is a $(D \times \mathcal{D})$-matrix. Additionally, the transformation may project the feature space into a new subspace where certain model assumption, for instance a diagonal covariance modeling, are more suitable than in the original space.

A restriction of the LDA approach is that the data distributions have to share a common covariance matrix. The LDA transformation has been extended for heteroscedastic data and derived in a ML framework in [Kumar & Andreou 98]. Thus LDA is a special case of HDA.

The basic idea to formulate the feature space projections in a maximum likelihood framework is to extend $\boldsymbol{\Theta}_D$ to a full $(\mathcal{D} \times \mathcal{D})$ transformation matrix

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_D \\ \boldsymbol{\Theta}_{\mathcal{D}-D} \end{bmatrix} \tag{5.22}$$

where $\boldsymbol{\Theta}_D$ denotes the first D rows of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_{\mathcal{D}-D}$ the remaining $\mathcal{D} - D$ rows, both consisting of $\mathcal{D}$ columns. The basic assumption for the extension of Θ is that the dimensions above D carry no significant class separability information and are thus modeled by class independent parameters. The class [1] dependent mean vectors μ_s

[1] For notational simplicity, the HMM states are chosen as classes to be separated by the HDA/LDA, but many other classes are suitable. In fact, in the RWTH system the HMM states emerged as best choice in terms of word error rate [Zolnay 03].

and covariances Σ_s are extended by class independent components μ_0 and Σ_0:

$$\tilde{\mu}_s = \begin{bmatrix} \mu_{s1} \\ \vdots \\ \mu_{sD} \\ \mu_{01} \\ \vdots \\ \mu_{0(\mathcal{D}-D)} \end{bmatrix} = \begin{bmatrix} \mu_s \\ \mu_0 \end{bmatrix} \tag{5.23}$$

$$\tilde{\boldsymbol{\Sigma}}_s = \begin{bmatrix} \boldsymbol{\Sigma}_s & 0 \\ 0 & \boldsymbol{\Sigma}_0 \end{bmatrix} \tag{5.24}$$

where μ_0 is a $(\mathcal{D}-D)$ dimensional vector and Σ_0 a matrix of size $(\mathcal{D}-D)\times(\mathcal{D}-D)$. Again, the parameters $\tilde{\mu}_s, \tilde{\Sigma}_s$ and $\boldsymbol{\Theta}$ are estimated using an ML approach:

$$(\tilde{\mu}_s^{\text{opt}}, \tilde{\boldsymbol{\Sigma}}_s^{\text{opt}}, \boldsymbol{\Theta}^{\text{opt}}) = \underset{(\tilde{\mu}_s, \tilde{\boldsymbol{\Sigma}}_s, \boldsymbol{\Theta})}{\text{argmax}} \left\{ \sum_{t=1}^{T} \sum_{s'=1}^{S} \gamma_{s'}(t) \left[\log \left(\frac{|\det \boldsymbol{\Theta}|^2}{|\det \boldsymbol{\Sigma}_{s'}|} \right) - \left((\boldsymbol{\Theta}\tilde{x} - \tilde{\mu}_{s'})^\top \tilde{\boldsymbol{\Sigma}}_{s'}^{-1} (\boldsymbol{\Theta}\tilde{x} - \tilde{\mu}_{s'}) \right) \right] \right\} \tag{5.25}$$

The partition of $\boldsymbol{\Theta}$ from Eq. (5.22) results in the following estimates for the mean and covariances [Kumar & Andreou 98]:

$$\begin{aligned} \mu_s^{\text{opt}} &= \boldsymbol{\Theta}_D \, \bar{x}_s \quad s = 1, \ldots, S \\ \mu_0^{\text{opt}} &= \boldsymbol{\Theta}_{\mathcal{D}-D} \, \bar{x} \\ \boldsymbol{\Sigma}_s^{\text{opt}} &= \boldsymbol{\Theta}_D \, \boldsymbol{W}_s \, \boldsymbol{\Theta}_D^\top \quad s = 1, \ldots, S \\ \boldsymbol{\Sigma}_0^{\text{opt}} &= \boldsymbol{\Theta}_{\mathcal{D}-D} \, \boldsymbol{T} \, \boldsymbol{\Theta}_{\mathcal{D}-D}^\top \end{aligned}$$

with

$$\begin{aligned} \bar{x}_s &= \frac{1}{N_s} \sum_{n=1}^{N_s} x_n \quad \text{with } x_n \in s, \quad s = 1, \ldots, S \\ \bar{x} &= \frac{1}{N} \sum_{n=1}^{N} x_n \\ \boldsymbol{W}_s &= \frac{1}{N_s} \sum_{n=1}^{N_s} (x_n - \bar{x}_s)(x_n - \bar{x}_s)^\top \quad s = 1, \ldots, S \\ \boldsymbol{T} &= \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^\top \end{aligned}$$

The resulting optimization formula for $\boldsymbol{\Theta}^{\text{opt}}$ is given as

$$\begin{aligned}\boldsymbol{\Theta}^{\text{opt}} = \underset{\boldsymbol{\Theta}}{\text{argmax}} \Bigg\{ & -\frac{N}{2} \log |\det \left(\boldsymbol{\Theta}_{\mathcal{D}-D} \boldsymbol{T} \boldsymbol{\Theta}_{\mathcal{D}-D}^{\top} \right)| \\ & - \sum_{s=1}^{S} \frac{N_s}{2} \log |\det \left(\boldsymbol{\Theta}_D \boldsymbol{W}_s \boldsymbol{\Theta}_D^{\top} \right)| \\ & + N \log |\det \boldsymbol{\Theta}| \Bigg\} \end{aligned} \tag{5.26}$$

No closed-form solution of Eq. (5.26) has been found yet and the equation has to be optimized numerically. For further details the reader is referred to [Kumar & Andreou 98]. There is a close relationship between HDA and STC or EMLLT covariance modeling, especially for a full transformation (i.e. $D = \mathcal{D}$), which is subject of the following Section.

5.3.4 Relationship between STC, EMLLT and HDA

HDA and LDA transformations are typically formulated as transformation of the observation vector. By extending the dimensionality reducing transformation $\boldsymbol{\Theta}_D$ according to Eq. (5.22), it is possible to rewrite the probability distribution

$$p(\tilde{x}|\tilde{\mu}_s, \tilde{\boldsymbol{\Sigma}}_s; \boldsymbol{\Theta}) = \frac{|\det \boldsymbol{\Theta}|^2}{\sqrt{|\det 2\pi\tilde{\boldsymbol{\Sigma}}_s|}} \exp\left(-\frac{1}{2} (\boldsymbol{\Theta}\tilde{x} - \tilde{\mu}_s)^{\top} \tilde{\boldsymbol{\Sigma}}_s^{-1} (\boldsymbol{\Theta}\tilde{x} - \tilde{\mu}_s) \right) \tag{5.27}$$

$$= \frac{|\det \boldsymbol{\Theta}|^2}{\sqrt{|\det 2\pi\tilde{\boldsymbol{\Sigma}}_s|}} \exp\left(-\frac{1}{2} (\tilde{x} - \nu_s)^{\top} \left(\boldsymbol{\Theta}^{-1} \tilde{\boldsymbol{\Sigma}}_s \boldsymbol{\Theta}^{-\top} \right)^{-1} (\tilde{x} - \tilde{\nu}_s) \right) \tag{5.28}$$

with

$$\nu_s = \boldsymbol{\Theta}^{-1} \tilde{\mu}_s \tag{5.29}$$

Using this relationship it is easy to see that the ML optimization criterion Eq. (5.25) is again very similar to Eq. (5.4), just as discussed for the STC transformation. Moreover, a comparison of Eq. (5.28) and Eq. (5.18) reveals that STC is equivalent (for $\boldsymbol{H} = \boldsymbol{\Theta}^{-1}$) to an HDA without dimensionality reduction (i.e. $\mathcal{D} = D$) and diagonal covariances $\boldsymbol{\Sigma}_s = \boldsymbol{\Sigma}_s^{\text{diag}}$ in Eq. (5.24). Consequently, EMLLT is also equivalent to HDA for $\Delta = D$. For $\Delta > D$ there exits a close relationship between EMLLT and HDA without dimensionality reduction. Both approaches can be formulated as

transformation of the covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_s^{\mathrm{H}} = \boldsymbol{\Theta}^{-1}\,\boldsymbol{\Sigma}_s^{\mathrm{H}}\,\boldsymbol{\Theta}^{-\top} = \boldsymbol{H}^{\mathrm{H}}\boldsymbol{\Sigma}_s^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}\top} \quad (\boldsymbol{H}^{\mathrm{H}} = \boldsymbol{\Theta}^{-1}) \tag{5.30}$$

$$\hat{\boldsymbol{\Sigma}}_s^{\mathrm{E}} = \boldsymbol{H}^{\mathrm{E}}\,\boldsymbol{\Sigma}_s^{\mathrm{E}}\,\boldsymbol{H}^{\mathrm{E}\top} \tag{5.31}$$

$$\tag{5.32}$$

with

$$\begin{aligned} \boldsymbol{\Sigma}_s^{\mathrm{H}} &\in \mathbb{R}^{D\times D} \\ \boldsymbol{H}^{\mathrm{H}} &\in \mathbb{R}^{D\times D} \\ \boldsymbol{\Sigma}_s^{\mathrm{E}} &\in \mathbb{R}^{\Delta\times\Delta}\,, \quad \text{diagonal} \\ \boldsymbol{H}^{\mathrm{E}} &\in \mathbb{R}^{D\times\Delta} \end{aligned}$$

and $D \leq \Delta \leq D(D+1)/2$.
The superscripts $^{\mathrm{H}}$ and $^{\mathrm{E}}$ have been introduced to distinguish between the **E**MLLT and the **H**DA approach, respectively. The main difference between Eq. (5.30) and Eq. (5.31) is the number of free parameters for the transformation matrix $\boldsymbol{H}$, given an equal number of parameters for $\boldsymbol{\Sigma}_s$. The difference becomes clearer if the Eqs. (5.30) and (5.31) are rewritten component-wise and using the upper value $\Delta = D(D+1)/2$:

$$\left(\hat{\boldsymbol{\Sigma}}_s^{\mathrm{H}}\right)_{ij} = \sum_{k,l=1}^{D} h_{ik}^{\mathrm{H}}\,\boldsymbol{\Sigma}_{s\;kl}^{\mathrm{H}}\,h_{jl}^{\mathrm{H}} \tag{5.33}$$

$$= \sum_{d=1}^{D} \boldsymbol{\Sigma}_{s\;dd}^{\mathrm{H}}\,h_{dd}^{\mathrm{H}}\,h_{dd}^{\mathrm{H}} + 2\sum_{k<l=1}^{D} \boldsymbol{\Sigma}_{s\;kl}^{\mathrm{H}}\,h_{ik}^{\mathrm{H}}\,h_{jl}^{\mathrm{H}} \tag{5.34}$$

$$\left(\hat{\boldsymbol{\Sigma}}_s^{\mathrm{E}}\right)_{ij} = \sum_{d=1}^{D} \left(\sigma_s^{\mathrm{E}}\right)_d^2 h_{id}^{\mathrm{E}}\,h_{jd}^{\mathrm{E}} + \sum_{d=D+1}^{D(D+1)/2} \left(\sigma_s^{\mathrm{E}}\right)_d^2 h_{id}^{\mathrm{E}}\,h_{jd}^{\mathrm{E}} \tag{5.35}$$

Thus

$$\hat{\boldsymbol{\Sigma}}_s^{\mathrm{H}} = \sum_{d=1}^{D} \boldsymbol{\Sigma}_{s\;dd}^{\mathrm{H}}\,h_d^{\mathrm{H}}\,h_d^{\mathrm{H}\top} + \sum_{\substack{k,l=1\\k\neq l}}^{D} \boldsymbol{\Sigma}_{s\;kl}^{\mathrm{H}}\,h_k^{\mathrm{H}}\,h_l^{\mathrm{H}\top} \tag{5.36}$$

$$\hat{\boldsymbol{\Sigma}}_s^{\mathrm{E}} = \sum_{d=1}^{D} \left(\sigma_s^{\mathrm{E}}\right)_d^2 h_d^{\mathrm{E}}\,h_d^{\mathrm{E}\top} + \sum_{d=D+1}^{D(D+1)/2} \left(\sigma_s^{\mathrm{E}}\right)_d^2 h_d^{\mathrm{E}}\,h_d^{\mathrm{E}\top} \tag{5.37}$$

where h_k denotes the appropriate column vector of $\boldsymbol{H}$. A comparison of Eqs. (5.36) and (5.37) shows that both approaches HDA (without dimensionality reduction) and EMLLT are not only equivalent for $\Delta = D$, but also for $\Delta = D(D+1)/2$. $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}}$

is expressed in terms of rank-one basis matrices $h_\bullet h_\bullet^\top$ in Eqs. (5.36) and (5.37). The difference for $D < \Delta < (D+1)/2$ is as follows: In the HDA approach these basis matrices are made up from only D vectors as the outer product $h_k h_l^\top$, $1 < k, l < D$. Thus, the basis matrices are restricted to the subspace which is spanned by building all possible products $h_k h_l$. On the other hand, the basis matrices for the EMLLT approach are unrestricted. The difference occurs if the number of free parameters Δ in $\mathbf{\Sigma}^{\mathrm{H}}_{s\,ij}$ or $\mathbf{\Sigma}^{\mathrm{E}}_{s\,ij}$ is restricted to be less than $D(D+1)/2$, which is equal to restricting the number of basis matrices in the set $\{h_\bullet h_\bullet^\top\}$. Eqs. (5.36) and (5.37) show that the diagonal elements (the first sum in the respective equation) remain unaffected. For the HDA approach the off-diagonal elements have to be generated by rank-one matrices which are build by the same vectors used for the diagonal elements, but only in different combinations. In the EMLLT approach on the other hand, the rank-one matrices generating the off-diagonal elements may be completely independent of those generating the diagonal elements of $\mathbf{\Sigma}^{\mathrm{E}}_s$. Therefore the HDA approach represents a smaller amount of possible matrices $\hat{\mathbf{\Sigma}}^{\mathrm{H}}_s$ with Δ independent parameters because of the restricted freedom in building the basis matrices. In the EMLLT approach, the number of parameters for the basis matrices $h^{\mathrm{E}}_d\, h^{\mathrm{E}^\top}_d$ is larger compared to the HDA approach, given an equal number of parameters Δ.

In summary, EMLLT is equivalent to HDA without dimensionality reduction for $\Delta = D$ and $\Delta = D(D+1)/2$. For $D < \Delta < D(D+1)$ the degree of freedom in the choice of possible subspaces for the covariance matrices $\hat{\mathbf{\Sigma}}_s$ is larger in the EMLLT approach, thus allowing a more flexible modeling of the covariance matrices.

5.3.5 Summary

The speaker independent transformations can thus be classified as follows:

Table 5.2: Overview of speaker independent linear transformations based on Eq. (5.3). The subscript s denotes a dependency on the HMM states.

transformation	$\boldsymbol{H}$	$\boldsymbol{\Sigma}$	remark
MLLT	$\boldsymbol{H} \in \mathbb{R}^{D \times D}$	$\boldsymbol{\Sigma}_s^{\text{diag}} \in \mathbb{R}^{D \times D}$	
EMLLT	$\boldsymbol{H} \in \mathbb{R}^{D \times \Delta}$	$\boldsymbol{\Sigma}_s^{\text{diag}} \in \mathbb{R}^{\Delta \times \Delta}$	$D \leq \Delta \leq D(D+1)/2$
HDA	$\boldsymbol{H} \in \mathbb{R}^{D \times \mathcal{D}}$	$\begin{pmatrix} \boldsymbol{\Sigma}_s & 0 \\ 0 & \boldsymbol{\Sigma}_0 \end{pmatrix}$	$\boldsymbol{\Sigma}_s \in \mathbb{R}^{D \times D}$ $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{(\mathcal{D}-D) \times (\mathcal{D}-D)}$ $D < \mathcal{D}$
LDA	$\boldsymbol{H} \in \mathbb{R}^{D \times \mathcal{D}}$	$\begin{pmatrix} \boldsymbol{\Sigma}^{\text{diag}} & 0 \\ 0 & \boldsymbol{\Sigma}_0 \end{pmatrix}$	$\boldsymbol{\Sigma}^{\text{diag}}$ globally pooled

- LDA is a special case of HDA
- STC is equivalent to HDA without dimensionality reduction and diagonal covariance modeling
- EMLLT is equivalent to HDA without dimensionality reduction and diagonal covariance modeling for for $\Delta = D$ and $\Delta = D(D+1)/2$, but allows for a more flexible covariance modeling for $D < \Delta < D(D+1)$
- HDA, STC and EMLLT are closely related to variance adaptation (c.f. Section 5.2.2)

5.4 Conclusion

In this Chapter the numerous linear transformations that are used in today's state-of-the-art speech recognition system have been presented in a unified view. While the original publications addressed the details of each individual approach, this Chapter has highlighted the strong interrelationship between the approaches.

Although motivated differently by the original authors, the linear transformations presented in the last Sections can be described in a unified way: All those

transformations can be derived from Eq. (5.3).

$$p(x|\mu_s, \boldsymbol{\Sigma}_s; \mathbf{A}, b, \boldsymbol{H}) = \frac{1}{\sqrt{|\det(2\pi \boldsymbol{H}\,\boldsymbol{\Sigma}_s\,\boldsymbol{H}^\top)|}} \exp\left(-\frac{1}{2}(x - \boldsymbol{A}\,\mu_s - b)^\top \left(\boldsymbol{H}\,\boldsymbol{\Sigma_s}\,\boldsymbol{H}^\top\right)^{-1}(x - \boldsymbol{A}\,\mu_s - b)\right) \tag{5.3}$$

The transformations are estimated using the maximum likelihood optimization function Eq. (5.4):

$$(\boldsymbol{A}^{\text{opt}}, b^{\text{opt}}, \boldsymbol{H}^{\text{opt}}) = \underset{(\boldsymbol{A},b,\boldsymbol{H})}{\text{argmax}} \left\{ \sum_{t=1}^{T} \sum_{s=1}^{S} \gamma_s(t) \left[\log\left(\frac{1}{|\det \boldsymbol{H}\,\boldsymbol{\Sigma}_s\,\boldsymbol{H}^\top|}\right) - \left((x_t - \boldsymbol{A}\,\mu_s - b)^\top \left(\boldsymbol{H}\,\boldsymbol{\Sigma}_s\,\boldsymbol{H}^\top\right)^{-1}(x_t - \boldsymbol{A}\,\mu_s - b)\right)\right]\right\} \tag{5.38}$$

The differences in the approaches result from different realizations of the transformation matrices, different modelings of the covariance matrix or the use of different data to estimate the transformation. For example, STC is closely related to variance adaptation: for STC the sum over t in Eq. (5.38) includes all time frames of the training data, whereas for variance adaptation the sum over t includes only the adaptation data[2] from one specific speaker. Close relationships have also been shown between STC, EMLLT and HDA. Another strong correlation exists between C-MLLR and VTN. As VTN can be expressed as linear transformation of the feature vector in the Cepstral domain (c.f. Chapter 7), VTN is in fact a C-MLLR with a highly restricted specific transformation matrix.

[2]or test data in case unsupervised adaptation

Chapter 6

Improved Signal Analysis

The signal analysis front end which is used in most of today's automatic speech recognition systems use some sort of either Mel-frequency Cepstral coefficients (MFCCs) [Davis & Mermelstein 80] or perceptual linear predictive (PLP) coding [Hermansky 90] signal analysis front end. A typical MFCC signal analysis front end was described in Section 1.2. The current Section focuses on the steps between the Fourier transform and the Cepstrum transform which are depicted in Fig 6.1, including vocal tract normalization (VTN) (c.f. Section 4.2.2). After some preprocessing

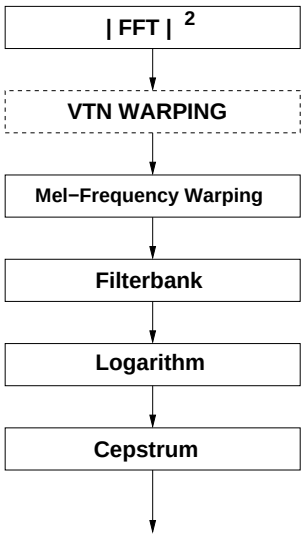

Figure 6.1: Typical MFCC signal analysis front end.

like preemphasis and windowing, the Fourier power spectrum of the speech waveform is computed for each time frame. The frequency axis of this power spectrum is warped by a warping function. When a VTN approach is used, the frequency axis is usually warped using a piece-wise linear or bilinear warping function with a speaker dependent warping factor which adjusts the amount of frequency warping. A Mel frequency warping [Davis & Mermelstein 80] is applied to adjust the spectral resolution to that of the human ear. Afterwards a filter bank is applied and the

logarithm is taken. In a last step a Cepstrum transformation, which is identical to a discrete cosine transformation for symmetric functions like the power spectrum, is applied to the log filter bank coefficients to remove the correlation between the different outputs. The dimensionality of the Cepstral vector is reduced by omitting the highest Cepstral coefficients for smoothing.

Mel frequency warping and VTN frequency warping are quite similar from the signal analysis point of view. They differ mainly by the specific choice of the warping function, which is in addition speaker dependent in case of VTN. Thus in the following paragraph only Mel frequency warping will be considered, but the statements hold for VTN as well. Frequency warping followed by a filter bank may be implemented in two different ways, which are shown in Fig. 6.2. One method is to explic-

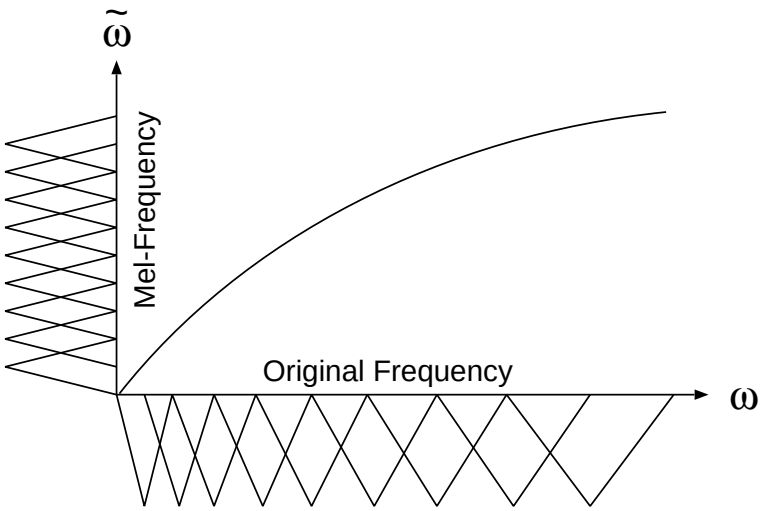

Figure 6.2: Basic principle of different filter bank implementations. The filters may either be uniformly distributed in the Mel-frequency domain (ordinate) or non-uniformly distributed in the original frequency domain (abscissa).

itly warp the frequency axis of the magnitude spectrum and apply uniformly shaped and spaced filters to the explicitly warped spectrum [Wegmann & McAllaster$^+$ 96]. A disadvantage of this approach is that the Mel-warped spectrum has to be obtained by interpolating from the original spectrum. This may introduce interpolation errors due to the large dynamic range of the magnitude spectrum. Another approach is to apply non-uniformly placed and shaped filters to the original magnitude spectrum [Lee & Rose 96]. This approach avoids explicitly warping the frequency axis and applies the Mel-warping implicitly. However, this approach is vulnerable to quantization errors if the spectral resolution is not appropriate. The lowest filter may cover a few spectral lines only and the maximum of a filter may fall between two spectral lines. In addition, this approach is problematic if an additional frequency warping is applied, for instance when using VTN [Lee & Rose 96, Chu & Jie$^+$ 97].

Experiments at the RWTH have also shown that in combination with VTN the first approach is superior.

The Cepstrum transformation is used to smooth the signal by omitting the higher Cepstral coefficients, which mainly describe the pitch. Due to the overlapping filters the different filter channels are correlated which leads to a covariance matrix of approximately Toeplitz structure. The Cepstrum transformation is applied to decorrelate the filter bank coefficients [Davis & Mermelstein 80]. This approach has two main disadvantages: Firstly, the optimal number and shape of the filter bank is unknown and is subject to optimization. For instance, experiments at RWTH have shown cosine shaped filters to outperform triangular filters for certain tasks. Secondly, this approach provides a twofold smoothing: one by applying the filter bank and the other by reducing the number of Cepstral coefficients. In general, each smoothing has to be optimized independently, i.e. the number of filters and the number of Cepstral coefficients. In the following a different approach will be described in which the frequency warping is integrated directly into the Cepstrum transformation and the filter bank is omitted completely [Molau & Pitz$^+$ 01]. Thus a multiple smoothing is avoided and the number of parameters to be optimized is reduced significantly.

6.1 Integrated Frequency Axis Warping

This Section presents an approach to calculate the Cepstral coefficients directly from the magnitude spectrum while omitting the usual filter bank. The Cepstral coefficients $c_k, k = 0, \ldots, K$ of an unwarped spectrum $X(\omega)$ are given by

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega \; \mathrm{e}^{i\omega k} \ln |X(\omega)|^2 \;, \quad k = 0, \ldots, K \tag{6.1}$$

The application of an invertible frequency axis warping function $g : [-\pi, \pi] \rightarrow [-\pi, \pi]$ can be expressed as follows:

$$\begin{aligned} g : [-\pi, \pi] &\rightarrow [-\pi, \pi] \\ \omega &\rightarrow \tilde{\omega} = g(\omega) && (6.2) \\ |\{X(\omega)\}| &= \left|\left\{\tilde{X}(g(\omega))\right\}\right| && (6.3) \\ &= \left|\left\{\tilde{X}(\tilde{\omega})\right\}\right| && (6.4) \end{aligned}$$

That means, for the new frequency $\tilde{\omega}$ the new spectrum $\tilde{X}(\tilde{\omega})$ is computed according to (see also Fig. 6.3):

$$\left|\tilde{X}(\tilde{\omega})\right| := \left|X\left(g^{(-1)}(\tilde{\omega})\right)\right|$$

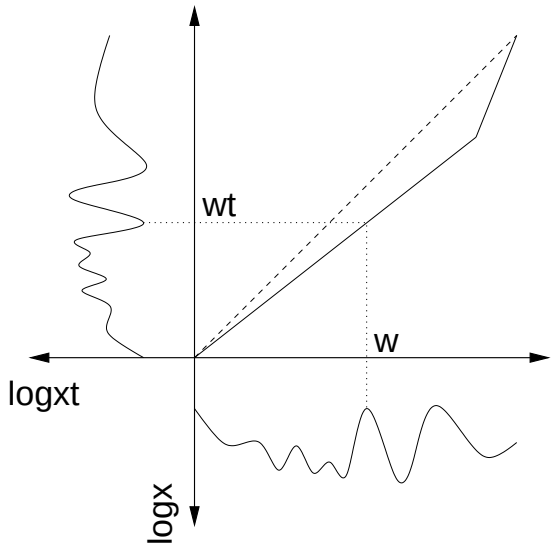

Figure 6.3: Basic principle of frequency warping using a piece-wise linear warping function.

Hence, the $k-$th Cepstral coefficient $\tilde{c}_k$ of the warped spectrum is given by

$$\tilde{c}_k = \frac{1}{\pi} \int_{-\pi}^{\pi} d\tilde{\omega}\, \mathrm{e}^{i\tilde{\omega}k} \ln \left| \tilde{X}(\tilde{\omega}) \right|^2 \tag{6.5}$$

$$= \frac{1}{\pi} \int_{-\pi}^{\pi} d\tilde{\omega}\, \mathrm{e}^{i\tilde{\omega}k} \ln \left| X\left(g^{(-1)}(\tilde{\omega})\right) \right|^2 \ . \tag{6.6}$$

Up to now, the frequency axis warping is applied to the original spectrum, expressed by the term $X\left(g^{(-1)}(\tilde{\omega})\right)$ in Eq. (6.6). In order to integrate the frequency warping into the Cepstrum transformation, the integration variable is changed from $\tilde{\omega}$ to ω in the usual way with taking the Jacobian determinant $d\tilde{\omega}/d\omega$ of the substitution into account:

$$\tilde{c}_k = \frac{1}{\pi} \int_{-\pi}^{\pi} d\omega\, \mathrm{e}^{ig(\omega)k} \ln |X(\omega)|^2 \cdot g'(\omega) \tag{6.7}$$

The boundaries of the integration do not change because of the constraints $\tilde{\omega}(-\pi) = \omega(-\pi)$ and $\tilde{\omega}(\pi) = \omega(\pi)$, c.f. Eq. (6.2).

Practical applications work with discrete spectra, which are band limited by the Nyquist frequency. Thus Eq. (6.7) is equivalent to

$$\tilde{c}_k = \frac{1}{N} \sum_{n=0}^{N/2-1} \left\{ \cos\left[g\left(\frac{2\pi n}{N}\right) k \right] \ln \left| X\left(\frac{2\pi n}{N}\right) \right|^2 \cdot g'\left(\frac{2\pi n}{N}\right) \right\} \quad , \tag{6.8}$$

where N is the FFT length. Equation (6.8) describes a matrix multiplication, which allows for a very compact implementation of the signal analysis. A schematic comparison of the traditional MFCC signal analysis with the presented integrated approach is depicted in Fig. 6.4.

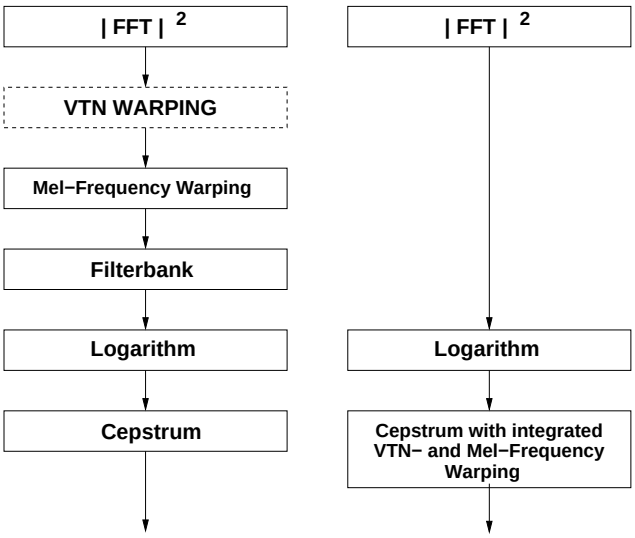

Figure 6.4: Schemes of traditional MFCC computation (left) and integrated approach (right).

In the following two sections the equations for Mel frequency warping and VTN using a piece-wise linear warping function will be derived explicitly.

6.1.1 Mel Frequency Warping

Mel frequency warping is commonly applied according to [Young 93]

$$\begin{aligned} \omega \rightarrow \tilde{\omega} &= m(\omega) \\ &= 2595 \cdot \log\left(1 + \frac{\omega f_s}{2\pi \cdot 700\text{Hz}}\right) \end{aligned} \tag{6.9}$$

where f_s denotes the sampling frequency in Hertz. The Mel frequency scale is introduced to adjust the spectral resolution to that of the human ear and the parameters 2595 and 700 were adjusted empirically by the authors of [Young 93]. In order to integrate the Mel frequency warping into the Cepstrum transform, Eq. (6.9) has to be normalized to meet the co-domain specification as given in Eq. (6.2). Thus the

following Mel frequency warping function is used:

$$\begin{aligned} g_{\text{mel}}(\omega) &= \pi \frac{m(\omega)}{m(\pi)} \\ &= d \cdot \log\left(1 + \frac{\omega f_s}{2\pi \cdot 700\text{Hz}}\right) \end{aligned} \tag{6.10}$$

with

$$d = \log\left(1 + \frac{f_s}{2\pi \cdot 700\text{Hz}}\right) \tag{6.11}$$

and the derivative needed in Eq. (6.8)

$$g'_{\text{mel}}(\omega) = \frac{d \cdot f_s}{(2\pi \cdot 700\text{Hz} + \omega \cdot f_s)\ln(10)} \tag{6.12}$$

Hence, substituting g_{mel} and g'_{mel} from Eq. (6.10) and Eq. (6.12), respectively, for g and g' in Eq. (6.8) yields the equation to compute the integrated Mel frequency warped Cepstral coefficients.

6.1.2 VTN Frequency warping

Vocal tract normalization (VTN) warps the frequency axis of the magnitude spectrum to reduce speaker dependency from the speech signal (c.f. Section 4.2). Usual implementations use either a modified filter bank as shown in Fig. 6.2 for Mel frequency warping [Lee & Rose 96] or an explicit warping of the frequency axis by interpolating the original discrete magnitude spectrum [Wegmann & McAllaster$^+$ 96, Welling 99]. As said before, VTN frequency warping and Mel frequency are quite similar from the signal analysis point of view. Hence, VTN frequency warping can equally well be integrated into the Cepstrum transformation. In the following this will be formulated exemplary for a piece-wise linear transformation function as shown in Fig. 6.5 . The formulation for other warping functions is similar. In order to simplify the notation and to avoid complicated case distinctions the warping function is written in the following form:

$$g_\alpha(\omega) = \beta_\omega\, \omega + \gamma_\omega \tag{6.13}$$

The parameters β_ω and γ_ω depend on ω only via the different sections of the piecewise linear function and can take only two discrete values as function of ω:

$$\beta_\omega = \begin{cases} \alpha & \omega \le \omega_0 \\ \frac{\pi - \alpha\omega_0}{\pi - \omega_0} & \omega > \omega_0 \end{cases} \tag{6.14}$$

$$\gamma_\omega = \begin{cases} 0 & \omega \le \omega_0 \\ (\alpha - 1)\frac{\pi\,\omega_0}{\pi - \omega_0} & \omega > \omega_0 \end{cases} \tag{6.15}$$

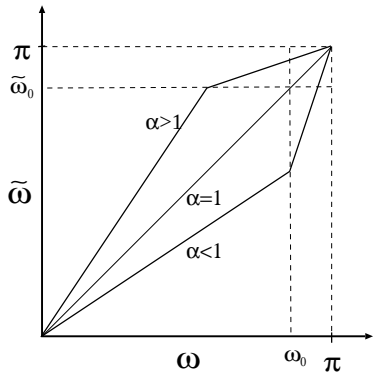

Figure 6.5: Schematic piece-wise linear VTN warping function.

The inflexion point ω_0, where the slope of the warping function changes, is chosen as follows:

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

Mel frequency warping is usually applied after VTN frequency warping and thus the combination of both frequency warping steps is given as

$$\begin{aligned} g_{\alpha,\mathrm{mel}}(\omega) &= g_{\mathrm{mel}}\left(g_\alpha(\omega)\right) \\ &= d \cdot \log\left(1 + \frac{(\beta_\omega\,\omega + \gamma_\omega) f_s}{2\pi \cdot 700\mathrm{Hz}}\right) \qquad (6.16) \\ g'_{\alpha,\mathrm{mel}}(\omega) &= \frac{d \cdot \beta_\omega \cdot f_s}{(2\pi \cdot 700\mathrm{Hz} + (\beta_\omega\,\omega + \gamma_\omega) f_s) \cdot \ln(10)}\,. \qquad (6.17) \end{aligned}$$

The Cepstral coefficients with combined Mel and VTN frequency warping using the presented integrated approach can now be calculated by substituting $g_{\alpha,\mathrm{mel}}$ and $g'_{\alpha,\mathrm{mel}}$ from Eqs. (6.16) and (6.17) for g and g' in Eq. (6.8), respectively.

6.2 Discussion

A comparison of the standard approach using a filter bank as described in Chapter 1.2 and the modified, integrated approach is shown in Fig. 6.6 for one sentence from the Verbmobil II corpus for the first and 15th Cepstral coefficient. While the first Cepstral coefficients are very similar for both approaches, the higher Cepstral coefficients differ significantly. The reason for the differences in the higher Cepstral coefficients lies in the additional smoothing provided by the filter bank. As the spectrum has been smoothed already by the filter bank in the traditional approach, the dynamic of the higher Cepstral coefficient is lower for the traditional MFCC

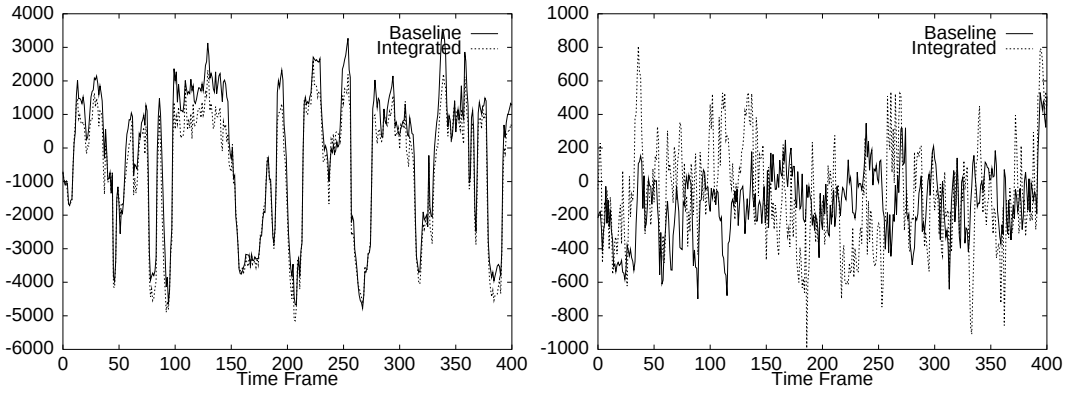

Figure 6.6: Comparison of Cepstrum coefficients 1 (left side) and 15 (right side) calculated using the traditional MFCC computation scheme (straight line) and the integrated approach (dashed line) for an utterance from the Verbmobil II corpus.

compared to the integrated MFCC. The multiple smoothing in the traditional approach leads to an interaction of both smoothing steps, which makes optimization of the parameters difficult. The integrated approach omits the filter bank and the speech signal is smoothed by reducing the number of Cepstral coefficients only. The advantage is that less parameters need to be optimized, which provides a better control of the smoothing step and avoids interaction of concurrent smoothing steps. Another advantage of the integrated approach is that the number of Cepstral coefficients, and thus the amount of smoothing, can be easily adjusted for a new task. For the traditional MFCC scheme, the number of MFCC components cannot be freely chosen because of the interaction with the number of filters in the filter bank. For instance, if a filter bank with 20 individual filters is used, the number of Cepstral coefficients may not exceed this number. A detailed discussion of this point is given in [Molau 03, pp. 80 ff.]

In addition, the mathematics to calculate the Cepstral coefficients are much simpler for the integrated approach than for the traditional MFCC. The filter bank makes analytic calculations, for example of frequency warping, more complicated. With the presented integrated approach it can be shown for instance that frequency warping of the Fourier spectrum amounts to a linear transformation of the Cepstral coefficients, which is subject of Chapter 7.

6.3 Experimental Evaluation

The improved signal analysis presented in this Chapter has been evaluated on two large vocabulary automatic speech recognition tasks, Verbmobil II (VM II) and the

North American Business News (NAB) task. The recognition tests were performed using the within-word RWTH ASR system described in Chapter 1, for details of the training and test corpora as well as the recognition test setup see Appendix C. The results are summarized in Table 6.1. The experimental setup is as follows:

- 20 filter bank channels (baseline approach only)
- 16 Mel frequency Cepstral coefficients
- NAB: sentence-wise long-term Cepstral mean normalization and energy normalization
  VM II: short-term Cepstral mean and variance normalization on a sliding window of 2 seconds length
- NAB: LDA on seven adjacent MFCC vectors, reduction to 32 dimensions
  VM II: LDA of three adjacent augmented MFCC vectors including first order derivatives, second order derivative of energy, reduction to 33 dimensions
- 3000 (NAB) / 2500 (VM II) decision-tree based generalized within-word triphone states plus one silence state
- 596k (NAB) /455k (VM II) Gaussian mixture densities
- 20000 word (NAB) / 10000 word (VMII) lexicon
- 6 state (NAB) / 3 state (VM II) HMM topology

Table 6.1: Recognition test results for the Verbmobil II and NAB tasks. Baseline: traditional MFCC computation scheme; integrated: new approach without filter bank and frequency warping integrated into the Cepstrum transformation [Molau & Pitz$^+$ 01].

Task	VTN	Cepstrum comp.	WER [%]
VM II	no	baseline	25.7
		integrated	25.3
	yes	baseline	23.8
		integrated	24.0
NAB	no	baseline	12.5
		integrated	12.4
	yes	baseline	11.8
		integrated	11.7

In general, the recognition performance of both the standard MFCC and the simpler and more compact integrated approach are quite similar. For the VM II baseline system the integrated approach achieves in addition a reduced word error rate.

6.4 Summary

In this Chapter an improved approach to compute Mel frequency Cepstral coefficients (MFCC) has been presented. In contrast to the traditional approach the frequency warping is integrated directly into the Cepstrum transform which avoids possible interpolation or quantization errors which may occur in the traditional approach. Any invertible frequency warping can easily be integrated in this approach. Explicit formula have been given for Mel frequency warping and a combination of Mel frequency warping and frequency warping in connection with vocal tract normalization (VTN). In addition, the new approach does not use any filter bank and thus avoids the twofold smoothing provided in the traditional approach by the filter bank and the subsequent reduction of Cepstral coefficients. The integrated approach allows for a very compact implementation and needs less parameters to be optimized for a new task. Experiments using the integrated approach have shown a similar recognition performance compared to the traditional approach.

Chapter 7

Frequency Warping as Linear Transformation in Cepstral Space

A speaker independent speech recognition system has to cope with a lot of variability in the acoustic signal. For example, varying transmission channels, noise, speakers, and speaking styles are sources of such irrelevant variabilities. From a more general perspective, this can be viewed as mismatch between training and testing condition of the ASR system. A lot of normalization (i.e. transformation of acoustic features) and adaptation (i.e. transformation of acoustic model parameters) schemes have been developed in the last years to compensate for this mismatch in order to improve the accuracy of the ASR system [Woodland 01].

7.1 Vocal Tract Normalization

A major part of the variability in the speech signal is caused by the speaker dependent vocal tract length. Vocal tract normalization (VTN) tries to compensate for the effect of speaker specific vocal tract lengths by warping the frequency axis of the power spectrum of the speech signal [Eide & Gish 96, Lee & Rose 96, Wakita 77, Wegmann & McAllaster$^+$ 96]. In a simple physiological model, the human vocal tract is treated as a uniform tube of length L. According to this model a change in L by a certain factor α^{-1} results in a scaling of the frequency axis by α. Thus, for this model, the frequency axis should be scaled linearly to compensate for the variability caused by different vocal tracts of individual speakers. However, the simple tube model may not be the best for the human vocal tract, so other frequency warping functions have been investigated [Eide & Gish 96]. In general, the frequency axis is scaled by a warping function g_a with a transformation parameter α

$$\begin{aligned} g_\alpha : [0,\pi] &\rightarrow [0,\pi] \\ \omega &\rightarrow \tilde{\omega} = g_\alpha(\omega) \end{aligned} \tag{7.1}$$

where ω denotes the original frequency and $\tilde{\omega}$ the warped frequency. The warping function g_α is assumed to be invertible, i.e. strictly monotonic and continuous

(see Fig. 7.1). The frequency $\omega = \pi$ corresponds to the Nyquist frequency and the domain and co-domain are chosen to conserve bandwidth and information contained in the original spectrum.

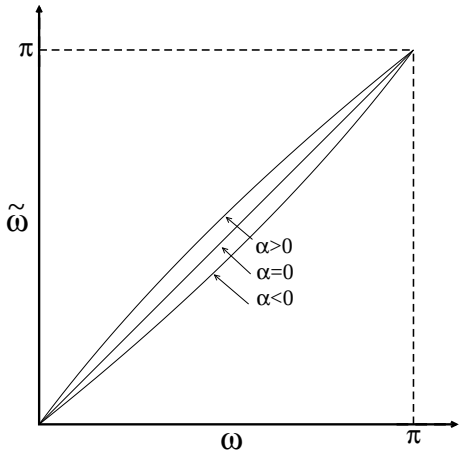

Figure 7.1: Example of a VTN warping function $\tilde{\omega} = g_\alpha(\omega)$ for different values of α.

All typical VTN approaches have in common that the warping function depends only on a few free parameters that control the amount of the frequency distortion. Even with only one free parameter (the warping factor α) to model the mismatch, VTN performs very well in a variety of recognition tasks. On the other hand, a few parameters allow to be estimated reliably on very little data, which makes VTN a good choice for on-line speech recognition systems.

The relationship between VTN and linear transformations in the Cepstral domain has been studied before. However, these investigations were restricted to a bilinear warping function [Acero 90, p.119],[McDonough 00, p.113], conformal mappings [McDonough & Zavaliagkos$^+$ 96], or were based on plausibility arguments and rough approximations [Cox 00]. In [Uebel & Woodland 99] it was shown that improvements in recognition accuracy gained by VTN and subsequent linear transformation (constrained MLLR) were not additive, which indicates that both techniques may not be independent of each other.

In this Chapter it will be shown for arbitrary invertible warping functions without approximations that there is a strong interdependence of VTN and a linear transformation of the Cepstral vector. A related result has also been reported in [Nocerino & Rabiner$^+$ 85] in the context of spectral distortion measures. For the case of Gaussian emission probabilities the equivalence of two widely used normalization approaches (namely VTN and MLLR) will be shown, which have so far been considered to be independent.

7.2 VTN Equals Linear transformation in Cepstral space

In the remaining of this Chapter the integrated MFCC computation scheme introduced in Chapter 6 will be used, in which the frequency warping is directly integrated into the Cepstrum transformation and the filter banks are omitted. In order to keep the equations simple, the investigation will be at first made for plain Cepstral coefficients (CC). Mel-scale warping will be considered later because the occurring integrals and thus the transformation matrix may be computed analytically for plain CC. In Section 7.5 it will be shown that the results of the current Section are still valid for MFCC.

The Cepstral coefficients c_k, $k = 0, \ldots, K$ of a spectrum $X(\omega)$ are defined by

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega \; \mathrm{e}^{i\omega k} \ln |X(\omega)|^2 \;, \quad k = 0, \ldots, K \\ &= \frac{s_k}{\pi} \int_{0}^{\pi} d\omega \; \cos(\omega k) \ln |X(\omega)|^2 \;, \end{aligned} \tag{7.2}$$

where ω may either denote the true physical or the Mel frequency scale. The symmetry factor s_k is introduced for convenience:

$$s_k = \begin{cases} \frac{1}{2} & : \quad k = 0 \\ 1 & : \quad \text{else} \; . \end{cases} \tag{7.3}$$

The $n-$th cepstral coefficient $\tilde{c}_n(\alpha)$ of the *warped* spectrum is given by

$$\tilde{c}_n(\alpha) = \frac{s_k}{\pi} \int_{0}^{\pi} d\tilde{\omega} \; \cos(\tilde{\omega} n) \ln \left| X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2 \;, \quad n = 1, \ldots, N \; . \tag{7.4}$$

The spectrum $\ln \left| X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2$ is expanded in a Fourier series

$$\begin{aligned} \ln \left| X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2 &= \ln |X(\omega)|^2 \\ &= 2 \sum_{k=0}^{K} c_k \cos(\omega k) \\ &= 2 \sum_{k=0}^{K} c_k \cos(g_\alpha^{(-1)}(\tilde{\omega}) k) \;, \end{aligned} \tag{7.5}$$

where c_k denotes the k-th Cepstral coefficient of the *unwarped* spectrum. In the case of continuous spectra, there may be no upper limit for K. Therefore it is

assumed that the original spectrum can be represented by a finite number of Cepstral coefficients, for instance if it has been Cepstrally smoothed already. This assumption is not mandatory as long as the Fourier series is uniformly convergent, which will be the case for the spectra dealt with in practical applications. Practical applications, however, work with discrete spectra. Hence, K will be finite and equal to the number of spectral lines of the discrete Fourier spectrum

In order to derive the transformation of the Cepstral coefficients, Eq. (7.5) is inserted into Eq. (7.2) and the integration and summation is interchanged. Thus, the warped Cepstral coefficients $\tilde{c}_n(\alpha), n = 1, \ldots, N$ are given as

$$\begin{aligned}
\tilde{c}_n(\alpha) &= \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}\ \cos(\tilde{\omega}n) \sum_{k=0}^{K} c_k \cos(g_\alpha^{(-1)}(\tilde{\omega})k) \\
&= \sum_{k=0}^{K} c_k\ \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}\ \cos(\tilde{\omega}n) \cos(g_\alpha^{(-1)}(\tilde{\omega})k) \\
&= \sum_{k=0}^{K} A_{nk}(\alpha)\ c_k
\end{aligned} \tag{7.6}$$

with

$$A_{nk}(\alpha) = \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}\ \cos(\tilde{\omega}n) \cos(g_\alpha^{(-1)}(\tilde{\omega})k) \tag{7.7}$$

with the symmetry factor from Eq. (7.3)

$$s_k = \begin{cases} \frac{1}{2} & : \quad k = 0 \\ 1 & : \quad \text{else} \ . \end{cases}$$

Eq. (7.6) describes a linear transformation of the original Cepstral coefficients c_k with a transformation matrix $\mathbf{A}(\alpha)$ of dimension $N \times K$ to obtain the vector of warped Cepstral coefficients $\tilde{c}_n(\alpha)$. In other words, VTN warping with a warping function g_α is expressed as a linear transformation of the Cepstral coefficients of the unwarped spectrum with transformation matrix $\mathbf{A}(\alpha)$ as given in Eq. (7.7). Note that the only assumption made concerning the warping function is invertibility.

For Cepstral smoothing the number of Cepstral coefficients may be reduced by omitting higher coefficients. In order to prevent a loss in spectral resolution the dimensionality reduction has to be applied to the *warped* Cepstral coefficients. This will be discussed in more detail in Section 7.6.

7.3 Analytic Calculation of the Transformation Matrix

As shown in the last Section, VTN can always be expressed as a linear transformation in the Cepstral domain, independent of the functional form of the (invertible) warping function (cf. Eqs. (7.6) and (7.7)). In the following the transformation matrix defined in Eq. (7.7) will be calculated analytically for 3 typical warping functions: piece-wise linear, quadratic and bilinear warping. The analytic calculation of the transformation matrix for other warping functions than those presented below, however, may not be as straightforward. Although the solutions for $\mathbf{A}(\alpha)$ will look quite different, it will be shown in Section 7.4 that the resulting matrices have a common shape. But first the detailed calculations will be presented.

7.3.1 Piece-wise Linear Warping Function

In order to apply a piece-wise linear warping, the solution for a strictly linear warping function is calculated first:

$$\begin{aligned} g_\alpha &: \omega \rightarrow \tilde{\omega} = \alpha \cdot \omega \\ g_\alpha^{(-1)} &: \tilde{\omega} \rightarrow \omega = \alpha^{-1} \cdot \tilde{\omega} \end{aligned}$$

This warping function does not meet the requirements of Eq. (7.1) as the co-domain is not equal to $[0, \pi]$ for $\alpha \neq 1$. This requirement will be neglected at first because the calculation for the piece-wise linear warping function (which meets the requirements of Eq. (7.1)) turns out to be very similar.

The entries $A_{nk}(\alpha)$ of the transformation matrix can be computed by elementary integration. For $\alpha \neq 1$ the solution is given as

$$\begin{aligned} A_{nk}(\alpha) &= \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega} \ \cos(\tilde{\omega} n) \cos(\alpha^{-1}\tilde{\omega} k) \\ &= \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \ \big(\cos(\tilde{\omega} n + \alpha^{-1}\tilde{\omega} k) + \cos(\tilde{\omega} n - \alpha^{-1}\tilde{\omega} k) \big) \\ &= s_k \frac{\sin[(n + \alpha^{-1}k)\pi]}{(n + \alpha^{-1}k)\pi} + s_k \frac{\sin[(n - \alpha^{-1}k)\pi]}{(n - \alpha^{-1}k)\pi} \ . \end{aligned}$$

For $\alpha = 1$ this simplifies to

$$A_{nk}(1) = \delta_{nk}$$

because of the orthonormality of the cosine function.

To meet the requirement of invertibility, the warping function is extended to be piece-wise linear [Wegmann & McAllaster+ 96, Welling & Kanthak+ 99] as shown

in Fig. 7.2:

$$\omega \rightarrow \tilde{\omega} = g_\alpha(\omega) = \begin{cases} \alpha\omega & : \quad \omega \leq \omega_0 \\ \alpha\omega_0 + \dfrac{\pi - \alpha\omega_0}{\pi - \omega_0}(\omega - \omega_0) & : \quad \omega > \omega_0 \end{cases} \tag{7.8}$$

The inflexion point ω_0, where the slope of the warping function changes, is chosen as follows:

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

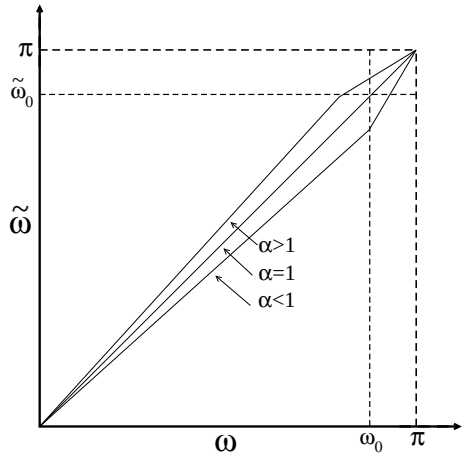

Figure 7.2: Piece-wise linear warping functions for $\alpha = 0.9, 1.0, 1.1$.

The transformation matrix $A_{nk}(\alpha)$ is computed similarly to the linear case but the integration is split into two parts:

$$A_{nk}(\alpha, \tilde{\omega}_0) = \frac{2s_k}{\pi} \left(\int_0^{\tilde{\omega}_0} + \int_{\tilde{\omega}_0}^{\pi} \right) d\tilde{\omega} \ \cos(\tilde{\omega}n) \cos(g_\alpha^{(-1)}(\tilde{\omega})k)$$

with $\tilde{\omega}_0 = \alpha \cdot \omega_0$.

Noting that the solution for $\alpha = 1$ remains the same as in the linear case, the solution for $\alpha \neq 1$ yields:

$$\begin{aligned} A_{nk}(\alpha) = & s_k \frac{\sin[(n - \alpha^{-1}k)\tilde{\omega}_0]}{(n - \alpha^{-1}k)\pi} + s_k \frac{\sin[(n + \alpha^{-1}k)\tilde{\omega}_0]}{(n + \alpha^{-1}k)\pi} \\ & - s_k \frac{\sin[(n - \alpha^{-1}k)\tilde{\omega}_0]}{\left(n - \frac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}k\right)\pi} - s_k \frac{\sin[(n + \alpha^{-1}k)\tilde{\omega}_0]}{\left(n + \frac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}k\right)\pi} \end{aligned} \tag{7.9}$$

This matrix can now be used for VTN alternatively to the conventional approaches like explicitly warping the power spectrum during signal analysis or the integrated approach described in Chapter 6. A detailed discussion of this warping matrix will be given in Section 7.4.

7.3.2 Quadratic warping function

In order to study the effect of the functional form of the warping function on the transformation matrix, the matrix is calculated for a quadratic warping function (Fig. 7.3). Although it has not been studied in literature in such a detail like piecewise or bilinear warping functions, a quadratic warping function is the next step beyond linear warping in a sense of a power series expansion.
The quadratic warping function is defined as follows:

$$g_\alpha^{(-1)} \quad : \quad \tilde{\omega} \rightarrow \omega = \tilde{\omega} + \alpha \left(\frac{\tilde{\omega}}{\pi} - \left(\frac{\tilde{\omega}}{\pi} \right)^2 \right) \tag{7.10}$$

Starting point is again Eq. (7.7):

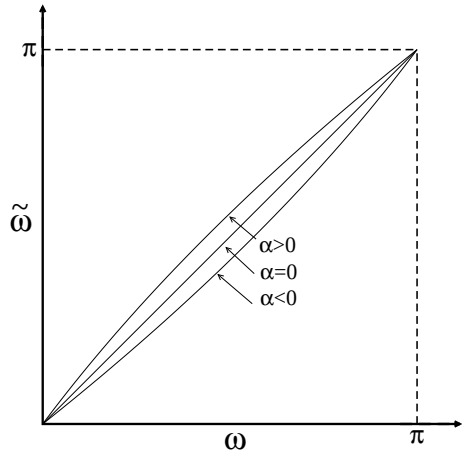

Figure 7.3: Example of quadratic VTN warping functions $\tilde{\omega} = g_\alpha(\omega)$ for $\alpha = -0.8,\ 0,\ +0.8$.

$$\begin{aligned} A_{nk}(\alpha) &= \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}\ \cos(\tilde{\omega}n) \cos(g_\alpha^{(-1)}(\tilde{\omega})k) \\ &= \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega}\ \left[\cos\left(\tilde{\omega}n + g_\alpha^{(-1)}(\tilde{\omega})k\right) + \cos\left(\tilde{\omega}n - g_\alpha^{(-1)}(\tilde{\omega})k\right) \right] \\ &= \frac{1}{\pi} \left(I_+ + I_-\right) \end{aligned} \tag{7.11}$$

with

$$
\begin{aligned}
I_{\pm} &= \int_0^{\pi} d\tilde{\omega}\ \cos\left\{\tilde{\omega}n \pm \left[\tilde{\omega} + \alpha\Big(\frac{\tilde{\omega}}{\pi} - (\frac{\tilde{\omega}}{\pi})^2\Big)\right]\ k\right\} \\
&= \int_0^{\pi} d\tilde{\omega}\ \cos\left\{\Big(n \pm k(1+\frac{\alpha}{\pi})\Big)\,\tilde{\omega} \mp k\frac{\alpha}{\pi^2}\tilde{\omega}^2\right\}\ .
\end{aligned}
\tag{7.12}
$$

The integrand of Eq. (7.12) has the form $\cos(ax^2 + bx)$ with $a = \mp k\alpha/\pi^2$ and $b = (n \pm k(1 + \alpha/\pi))$. To calculate $I_{\pm}$, the identity

$$
\cos(ax^2 + bx) = \cos(a(x + \frac{b}{2a})^2)\cos(\frac{b^2}{4a}) + \sin(a(x + \frac{b}{2a})^2)\sin(\frac{b^2}{4a})\ . \tag{7.13}
$$

is used. The resulting integrals from Eq. (7.13) can now be solved elementary with the solutions

$$
\begin{aligned}
\int dx\ \sin(a(x + \frac{b}{2a})^2) &= \sqrt{\frac{\pi}{2a}}\,\mathrm{S}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right) \\
\int dx\ \cos(a(x + \frac{b}{2a})^2) &= \sqrt{\frac{\pi}{2a}}\,\mathrm{C}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right)
\end{aligned}
\tag{7.14}
$$

where $\mathrm{S}(x)$ and $\mathrm{C}(x)$ denote the Fresnel sine and cosine which are defined as

$$
\begin{aligned}
\mathrm{S}(x) &= \int_0^x dt\ \sin(\frac{\pi}{2}\,t^2) \\
\mathrm{C}(x) &= \int_0^x dt\ \cos(\frac{\pi}{2}\,t^2)\ ,
\end{aligned}
$$

respectively. Thus, integration of Eq. (7.13) yields

$$
\begin{aligned}
&\int dx\ \cos(ax^2 + bx) = \\
&\sqrt{\frac{\pi}{2a}}\left[\cos(\frac{b^2}{4a})\,\mathrm{C}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right) + \sin(\frac{b^2}{4a})\,\mathrm{S}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right)\right]\ .
\end{aligned}
\tag{7.15}
$$

Finally, the solution of Eq. (7.11) for the quadratic warping function as defined

in Eq. (7.10) becomes:

$$
\begin{aligned}
&A_{nk}(\alpha) = \\
&\frac{\pi s_k}{\sqrt{2\pi k\alpha}} \cos\left(\frac{(n\pi + k(\pi+\alpha))^2}{4k\alpha}\right) \cdot \left[\mathrm{C}\left(\frac{n\pi + k(\pi+\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{C}\left(\frac{n\pi + k(\pi-\alpha)}{\sqrt{2\pi ka}}\right)\right] \\
+&\frac{\pi s_k}{\sqrt{2\pi k\alpha}} \cos\left(\frac{(n\pi - k(\pi+\alpha))^2}{4k\alpha}\right) \cdot \left[\mathrm{C}\left(\frac{n\pi - k(\pi-\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{C}\left(\frac{n\pi - k(\pi+\alpha)}{\sqrt{2\pi ka}}\right)\right] \\
+&\frac{\pi s_k}{\sqrt{2\pi k\alpha}} \sin\left(\frac{(n\pi + k(\pi+\alpha))^2}{4k\alpha}\right) \cdot \left[\mathrm{S}\left(\frac{n\pi + k(\pi+\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{S}\left(\frac{n\pi + k(\pi-\alpha)}{\sqrt{2\pi ka}}\right)\right] \\
+&\frac{\pi s_k}{\sqrt{2\pi k\alpha}} \sin\left(\frac{(n\pi - k(\pi+\alpha))^2}{4k\alpha}\right) \cdot \left[\mathrm{S}\left(\frac{n\pi - k(\pi-\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{S}\left(\frac{n\pi - k(\pi+\alpha)}{\sqrt{2\pi ka}}\right)\right]
\end{aligned}
\tag{7.16}
$$

This matrix will also be discussed in Section 7.4.

7.3.3 Bilinear warping function

Another frequently used warping function is the bilinear transformation (BLT). The calculation of the transformation matrix for a BLT is done in the complex z-domain instead of the real frequency ω because the occurring integrals will than be easier to compute. The BLT for a complex number z is defined by

$$\tilde{z} = \frac{z+\alpha}{1+\alpha z} \quad \text{with } z = \mathrm{e}^{i\omega},\ \tilde{z} = \mathrm{e}^{i\tilde{\omega}}$$

with the (real) warping parameter α, $|\alpha| < 1$. As the BLT has a simple form for complex numbers the z-transform is used to calculate the Cepstral coefficients, which is equivalent to the Fourier transform for $z = \mathrm{e}^{i\omega}$:

$$\log[X(z)] = \sum_{k=0}^{\infty} c_k z^{-k} \qquad c_k = \frac{1}{2\pi i} \oint dz\ \log[X(z)] z^{k-1}$$

The z-transform of warped spectrum is given by

$$\log[\hat{X}(\hat{z})] = \sum_{n=0}^{\infty} \hat{c}_n \hat{z}^{-n} \qquad \hat{c}_n = \frac{1}{2\pi i} \oint d\hat{z}\ \log[X(\hat{z})] \hat{z}^{n-1}\ .$$

Frequency warping is defined as

$$\hat{X}(\mathrm{e}^{i g_\alpha(\omega)}) = X(\mathrm{e}^{i\omega}) \quad \text{or} \quad \hat{X}(\hat{z}) = X(z)$$

and thus

$$\hat{c}_n = \frac{1}{2\pi i} \oint dz\ \sum_{k=0}^{\infty} c_k \hat{z}^{-k} z^{n-1}\ .$$

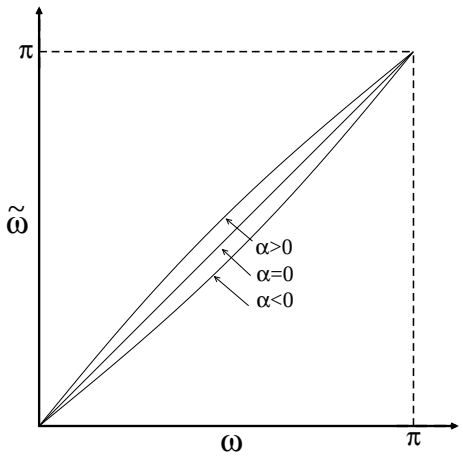

Figure 7.4: Example of bilinear VTN warping functions $\tilde{\omega} = g_\alpha(\omega)$ for $\alpha = -0.1,\ 0,\ 0.1$.

Interchanging integration and summation yields

$$\hat{c}_n = \sum_{k=0}^{\infty} \frac{1}{2\pi i} \oint dz\ \hat{z}^{-k} z^{n-1} \cdot c_k$$

Again, warping of the frequency axis of a given spectrum amounts to a linear transformation in the Cepstral domain

$$\hat{c}_n(\alpha) = \sum_k A_{nk}(\alpha)\, c_k$$

with the transformation matrix

$$A_{nk}(\alpha) = \frac{1}{2\pi i} \sum_{k=0}^{\infty} \oint dz\ \hat{z}^{-k} z^{n-1}\ .$$

The warping matrix $A_{nk}(\alpha)$ for a bilinear warping function has previously been calculated in [McDonough 00]:

$$A_{nk} = \frac{1}{2\pi i} \oint dz\ \left(\frac{z-\alpha}{1-\alpha z}\right)^{-k} z^{n-1}$$

The integration is carried out using the Cauchy integral formula

$$\frac{1}{n!} \left.\frac{d^{(n)} F(z)}{dz^{(n)}}\right|_{z=z_0} = \frac{1}{2\pi i} \oint d\xi\ \frac{F(\xi)}{(\xi - z_0)^{(n+1)}}$$

which yields

$$A_{nk}(\alpha) = \frac{1}{(k-1)!} \sum_{m=\max(0,k-n)}^{k} \binom{k}{m} \frac{(m+n-1)!}{(m+n-k)!} (-1)^m \alpha^{(2m+n-k)} \tag{7.17}$$

7.4 Discussion of the Structure of the Transformation Matrix

Having obtained the analytical solutions for the transformation matrix $A_{nk}(\alpha)$ in the last Section, the structure of the resulting matrices will be discussed. It will be shown that the matrices have a common structure even though the functional forms in Eqs. (7.9), (7.16) and (7.17) look quite different. Motivated by this common structure a possible approximation will be given to reduce the number of matrix elements to be calculated.

In the following, pictures of the transformation matrices are shown for values of the warping parameter α which occur typically in speech recognition experiments. As can be seen from Fig. 7.5 - 7.10, all matrices are dominated by the diagonal elements and the shapes of the matrices are similar.

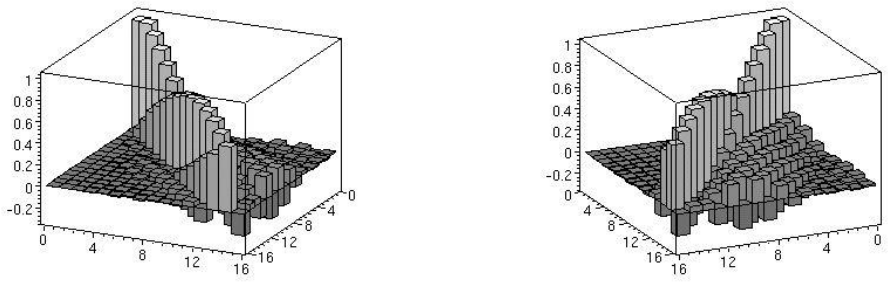

Figure 7.5: Matrix for piece-wise linear warping function, $\alpha = 0.9$.

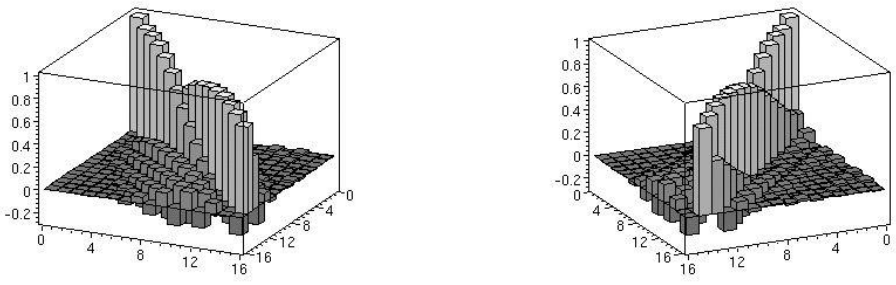

Figure 7.6: Matrix for piece-wise linear warping function, $\alpha = 1.1$.

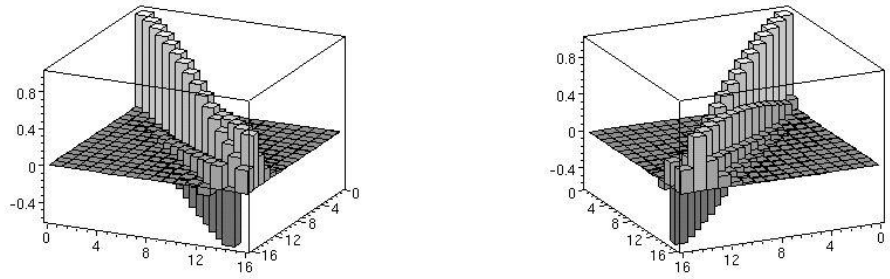

Figure 7.7: Matrix for quadratic warping function, $\alpha = -0.5$.

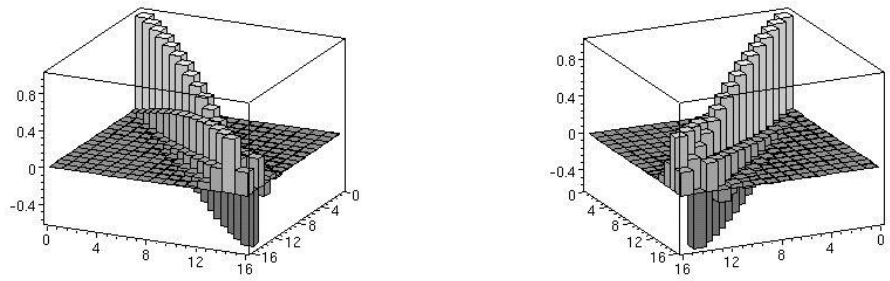

Figure 7.8: Matrix for quadratic warping function, $\alpha = +0.5$.

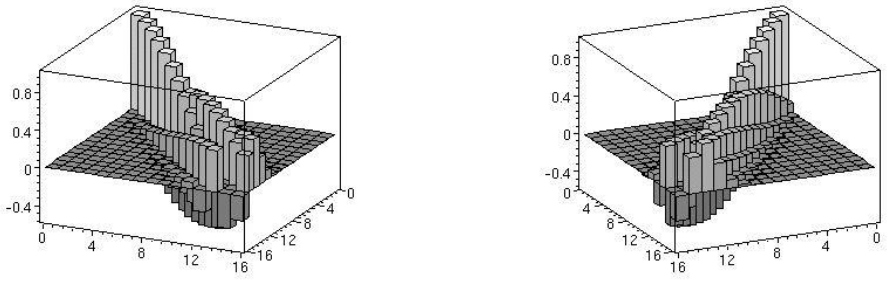

Figure 7.9: Matrix for bilinear warping function, $\alpha = +0.1$.

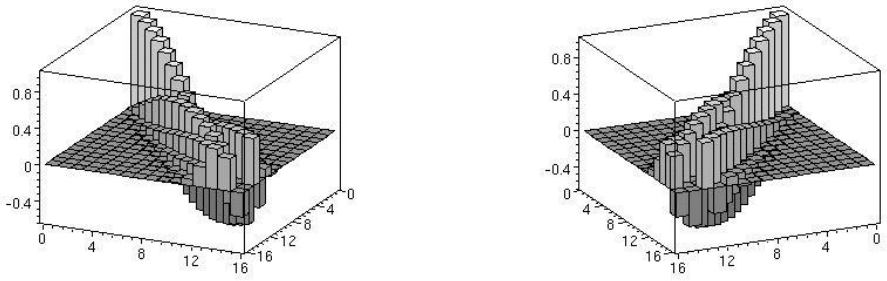

Figure 7.10: Matrix for bilinear warping function, $\alpha = -0.1$.

The dominance of the diagonal elements is a consequence of a general characteristic of typical warping functions rather than the actual functional form. This will lead to an approximation of the transformation matrix by a quindiagonal matrix.

Warping factors estimated for real speakers lead to warping functions which deviate only slightly from unity. This can also be seen from Fig. 7.2 to 7.4, where typical values of the warping factor α have been used. Hence the warping function can be written in the following form:

$$g_\alpha^{(-1)}(\tilde{\omega}) = \tilde{\omega} - \delta_\alpha(\tilde{\omega}) \quad \text{with} \quad \delta_\alpha(\tilde{\omega}) \ll 1 \; . \tag{7.18}$$

Further the deviation from unity $\delta_\alpha(\tilde{\omega}) := \tilde{\omega} - g_\alpha^{(-1)}(\tilde{\omega})$ is assumed to be smooth and to be either concave or convex.

Starting point of the following investigation is again Eq. (7.7):

$$\begin{aligned} A_{nk}(\alpha) &= \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega} \; \cos(\tilde{\omega} n) \cos(g_\alpha^{(-1)}(\tilde{\omega}) k) \\ &= \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \left[\cos\left(\tilde{\omega} n + g_\alpha^{(-1)}(\tilde{\omega}) k\right) + \cos\left(\tilde{\omega} n - g_\alpha^{(-1)}(\tilde{\omega}) k\right) \right] . \end{aligned} \tag{7.19}$$

Substituting Eq. (7.18) into Eq. (7.7) and using

$$\cos(x \pm y) = \cos(x)\cos(y) \mp \sin(x)\sin(y) \tag{7.20}$$

yields

$$\begin{aligned} A_{nk}(\alpha) = &\frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\ \cos[(n+k)\tilde{\omega}]\cos[k\,\delta_\alpha(\tilde{\omega})] \\ &-\frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\ \sin[(n+k)\tilde{\omega}]\sin[k\,\delta_\alpha(\tilde{\omega})] \\ &+\frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\ \cos[(n-k)\tilde{\omega}]\cos[k\,\delta_\alpha(\tilde{\omega})] \\ &+\frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\ \sin[(n-k)\tilde{\omega}]\sin[k\,\delta_\alpha(\tilde{\omega})] \end{aligned} \tag{7.21}$$

This equation is expanded for small $\delta_\alpha(\tilde{\omega})$ up to linear terms in $\delta_\alpha(\tilde{\omega})$ for $n \neq k$

$$\begin{aligned} A_{nk}(\alpha) = &\frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\ \cos[(n+k)\,\tilde{\omega}] - k\sin[(n+k)\,\tilde{\omega}]\,\delta_\alpha(\tilde{\omega}) \\ &+\frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\ \cos[(n-k)\,\tilde{\omega}] + k\sin[(n-k)\,\tilde{\omega}]\,\delta_\alpha(\tilde{\omega}) \\ &+\mathcal{O}(\delta_\alpha^2) \end{aligned} \tag{7.22}$$

It has been assumed that $\delta_\alpha(\tilde{\omega})$ is smooth and has a uniform curvature. Thus and because of $\delta_\alpha(\tilde{\omega}) \ll 1$, $\delta_\alpha(\tilde{\omega})$ can be considered as constant w.r.t. $\tilde{\omega}$ over each period

$$\left[(m-1)\frac{2\pi}{n\pm k}, m\frac{2\pi}{n\pm k}\right] \tag{7.23}$$

$(m = 1, \ldots, \left\lfloor \frac{n\pm k}{2} \right\rfloor)$ if the sine term oscillates quickly, i.e. for larger values of $(n+k)$ and $(n-k)$, respectively. Hence, $A_{nk}(\alpha)$ can be approximated by

$$A_{nk}(\alpha) \cong \delta_{nk} + \frac{k}{\pi}\int_0^\pi d\tilde{\omega}\ \sin[(n-k)\,\tilde{\omega}]\,\delta_\alpha(\tilde{\omega}) \tag{7.24}$$

where the second term contributes significantly only for small values of $(n-k), n \neq k$, i.e. close to the diagonal elements and becomes small for large values of $|n-k|$.

Thus the diagonal and only a few off-diagonal elements with small values of $|n-k|$ will dominate the matrix.

For $n = k$ the expansion has to be extended up to quadratic terms in $\delta_\alpha(\tilde{\omega})$ because the term linear in $\delta_\alpha(\tilde{\omega})$ vanishes, as can be seen from Eq. (7.24):

$$A_{nn} = 1 - s_n \frac{n}{2} \delta_\alpha^2(\tilde{\omega}) \tag{7.25}$$

Combining Eq. (7.24) and (7.25) yields

$$A_{nk}(\alpha) \cong \left(1 - \frac{s_n n}{2} \delta_\alpha^2(\tilde{\omega})\right) \delta_{nk} + \frac{k}{\pi} \int_0^\pi d\tilde{\omega}\ \sin[(n-k)\,\tilde{\omega}]\, \delta_\alpha(\tilde{\omega}) \tag{7.26}$$

Eq. 7.26 describes a matrix that is dominated strongly by the main and a few secondary diagonals. Fig. 7.5 - 7.10, which show typical warping functions obtained with characteristic warping factors, suggest an approximation by a quindiagonal matrix. Thus, for values of the warping factor α typical for automatic speech recognition, from Eq. 7.26 and Fig. 7.5 - 7.10 an approximation of the VTN warping matrix by a quindiagonal matrix is reasoned:

$$A_{nk}(\alpha) \cong 0 \quad \text{for} \quad |n-k| > 2\ . \tag{7.27}$$

It will be shown in the next Section that the matrix can be even approximated by a tridiagonal matrix when using the Mel frequency scale.

7.5 Integration of Mel Frequency Scale

In this Section it will be shown that Mel frequency scaling can equally well be integrated into the framework of VTN as linear transformation of the Cepstral Coefficients. Mel frequency warping is applied during signal analysis to adjust the spectral resolution to that of the human ear [Young 93]:

$$f_{mel} = 2595 \cdot \lg\left(1 + \frac{f}{700\text{Hz}}\right)\ . \tag{7.28}$$

So far, Mel frequency warping has not been considered. There are two possible ways to include Mel frequency warping into the framework of VTN as linear transformation of the Cepstral coefficients (CC):

A.) to express the VTN-Mel warped CC as a linear function of the original, unwarped CC or

B.) to express the VTN-Mel warped CC as a linear function of Mel-warped CC (MFCC) (without VTN).

Both methods will be discussed in the next two Subsections.

7.5.1 From Plain CC to VTN-Mel Warped CC

It has been shown in Section 7.2 that a frequency warping of the spectrum with an arbitrary invertible function results in a linear transformation of the Cepstral coefficients. Mel frequency warping can be considered as one special case of such a frequency warping and thus results in a linear transformation as well. Accordingly, the combination of VTN and subsequent Mel warping still amounts to a linear transformation in the Cepstral domain. VTN is typically applied before Mel scale warping; hence the combination of both warping steps becomes

$$g_{\text{mel}}(g_\alpha(\omega)) \;:\; \omega \rightarrow \tilde{\omega}_{\text{mel}} = B \cdot \lg\left(1 + \frac{g_\alpha(\omega) \cdot f_s}{2\pi \cdot 700\text{Hz}}\right) \tag{7.29}$$

where $g_\alpha(\omega)$ denotes the VTN warping function as before, f_s denotes the sampling frequency, and B is defined as

$$B = \frac{\pi}{\lg\left(1 + \frac{f_s}{2 \cdot 700Hz}\right)} \tag{7.30}$$

to meet the requirement $g_{\text{mel}}(\pi) = \pi$. Inserting Eq. (7.29) into Eq. (7.7) leads to

$$A_{nk}(\alpha) = \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}_{\text{mel}} \cos(\tilde{\omega}_{\text{mel}} n) \cos\left(g_\alpha^{(-1)}\left(g_{\text{mel}}^{(-1)}(\tilde{\omega}_{\text{mel}})\right) k\right) \tag{7.31}$$

Thus the Cepstral coefficients of the VTN-Mel-warped spectrum can be expressed as linear transformation of the original, unwarped Cepstral coefficients.

7.5.2 From MFCC to VTN Warped MFCC

It will be shown in Section 7.7 that VTN is equivalent to a parameterized constrained MLLR (C-MLLR) transformation. MLLR transforms the model parameters of the ASR, usually the means and/or covariances of the emission probability distributions, which have typically been estimated from Mel warped feature vectors. Thus more interesting and of practical relevance is to express the VTN-Mel-warped Cepstral coefficients as a function of the MFCC (i.e. without VTN) instead of the unwarped Cepstral coefficients. The difficulty in the present case is that VTN is typically applied *before* Mel warping. VTN-Mel-warped Cepstral coefficients $\tilde{c}_n^{\text{mel}}(\alpha)$ are defined as

$$\tilde{c}_n^{\text{mel}}(\alpha) = \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega}_{\text{mel}} \ln\left|\hat{X}(\tilde{\omega}_{\text{mel}})\right| \cos(\tilde{\omega}_{\text{mel}} n) \;. \tag{7.32}$$

VTN is usually applied to original, i.e. non-Mel-scaled, spectrum ($\tilde{\omega}_{\text{mel}}$ denotes the VTN-Mel-warped frequency). This frequency is given as

$$\tilde{\omega}_{\text{mel}} = \left(g_{\text{mel}} \circ g_\alpha\right)(\omega) \tag{7.33}$$

and the warped spectrum $\left\{\hat{X}(\tilde{\omega}_{\text{mel}})\right\}$ is obtained according to

$$\left\{\hat{X}(\tilde{\omega}_{\text{mel}})\right\} = \left\{X\left(g_{\alpha}^{(-1)}\left(g_{\text{mel}}^{(-1)}(\tilde{\omega}_{\text{mel}})\right)\right)\right\} = \left\{X(\omega)\right\} . \tag{7.34}$$

The log-spectrum is expanded as function of the Mel-warped frequency ω_{mel} in terms of unnormalized (i.e. not VTN-warped) Cepstral coefficients c_k^{mel}

$$\ln|X(\omega)|^2 = \ln\left|\hat{X}(\omega_{\text{mel}})\right|^2 = 2\sum_{k=0}^{K} c_k^{\text{mel}} \cos(\omega_{\text{mel}} k) . \tag{7.35}$$

As before, inserting Eq. (7.35) into Eq. (7.31) results in

$$\tilde{c}_n^{\text{mel}}(\alpha) = \sum_{k=0}^{K} c_k^{\text{mel}} \, \frac{2s_k}{\pi} \int_0^{\pi} d\tilde{\omega}_{\text{mel}} \;\; \cos(\omega_{\text{mel}} k) \cdot \cos(\tilde{\omega}_{\text{mel}} n) \tag{7.36}$$

The unnormalized Mel-scale frequency ω_{mel} needs to be expressed as function of the VTN-warped Mel-scale frequency $\tilde{\omega}_{\text{mel}}$:

$$\omega_{\text{mel}} = g_{\text{mel}}(\omega) = \left(g_{\text{mel}} \circ g_{\alpha}^{(-1)} \circ g_{\text{mel}}^{(-1)}\right)(\tilde{\omega}_{\text{mel}}) \tag{7.37}$$

and finally

$$\tilde{c}_n^{\text{mel}}(\alpha) = \sum_{k=0}^{K} A_{nk}^{\text{mel}}(\alpha) \; c_k^{\text{mel}} \tag{7.38}$$

with

$$A_{nk}^{\text{mel}}(\alpha) = \frac{2s_k}{\pi} \int_0^{\pi} d\tilde{\omega} \;\; \cos(\tilde{\omega} n) \cos\left(\left(g_{\text{mel}} \circ g_{\alpha}^{(-1)} \circ g_{\text{mel}}^{(-1)}\right)(\tilde{\omega}) \; k\right) . \tag{7.39}$$

Hence, the Cepstral coefficients $\tilde{c}_n^{\text{mel}}(\alpha)$ of the VTN-warped Mel-scale spectrum can be computed by a linear transformation of the unnormalized Cepstral coefficients c_k^{mel} (without VTN warping). Due to the highly non-linear transformation $g_{\text{mel}} \circ g_{\alpha}^{(-1)} \circ g_{\text{mel}}^{(-1)}$, the integral in Eq. (7.39) may not be solved analytically. Nevertheless, the transformation matrix can be calculated numerically.

The resulting warping function $g_{\text{eff}} := g_{\text{mel}} \circ g_{\alpha}^{(-1)} \circ g_{\text{mel}}^{(-1)}$ reads

$$g_{\text{eff}}(\tilde{\omega}_{\text{mel}}) := \left(g_{\text{mel}} \circ g_{\alpha}^{(-1)} \circ g_{\text{mel}}^{(-1)}\right)(\tilde{\omega}_{\text{mel}}) =$$

$$\begin{cases} B \cdot \log\left[1 + \frac{1}{\alpha}\left(10^{\tilde{\omega}_{\text{mel}}/B} - 1\right)\right] & : \quad \tilde{\omega}_{\text{mel}} \leq g_{\text{mel}}(\tilde{\omega}_0) \\ B \cdot \log\left[1 + \dfrac{f_s \tilde{\omega}_0}{2 \cdot 700\text{Hz}}\left(\dfrac{1}{\alpha} - \dfrac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}\right) + \dfrac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}\left(10^{\tilde{\omega}_{\text{mel}}/B} - 1\right)\right] & : \quad \tilde{\omega}_{\text{mel}} > g_{\text{mel}}(\tilde{\omega}_0) \end{cases} \tag{7.40}$$

In Fig. 7.11 the *inverse* warping function $\tilde{\omega}_{\text{mel}} = g_{\text{eff}}^{(-1)}(\omega)$ (straight line) is shown since this view is more familiar.

Expanding the effective warping function $g_{\text{eff}}(\tilde{\omega}_{\text{mel}})$ for $\tilde{\omega}_{\text{mel}} \leq g_{\text{mel}}(\tilde{\omega}_0)$ in a Taylor series around $\alpha = 1$

$$g_{\text{eff}}(\tilde{\omega}_{\text{mel}}) = \tilde{\omega}_{\text{mel}} - B\frac{1 - 10^{-\tilde{\omega}_{\text{mel}}/B}}{\ln(10)}(\alpha - 1) + \mathcal{O}\left((\alpha - 1)^2\right) \ . \tag{7.41}$$

(for $\tilde{\omega}_{\text{mel}} \geq g_{\text{mel}}(\tilde{\omega}_0)$ the expansion is similar) shows that the linear term dominates the expansion because the term $\frac{1-10^{-\tilde{\omega}_{\text{mel}}/B}}{\ln(10)}$ is small for $0 \leq \tilde{\omega}_{\text{mel}} \leq \pi$. Thus, $g_{\text{eff}}(\tilde{\omega}_{\text{mel}})$ may be approximated by a linear function with an appropriate choice of an effective warping factor α_{eff}. This matter will not be followed up in this work and will be subject of further research.

The Cepstral coefficients $\tilde{c}_n^{\text{mel}}(\alpha)$ obtained by the method presented here are identical to those calculated by explicitly warping the spectrum during signal analysis as presented in [Molau & Pitz$^+$ 01]. Comparing the resulting warping function

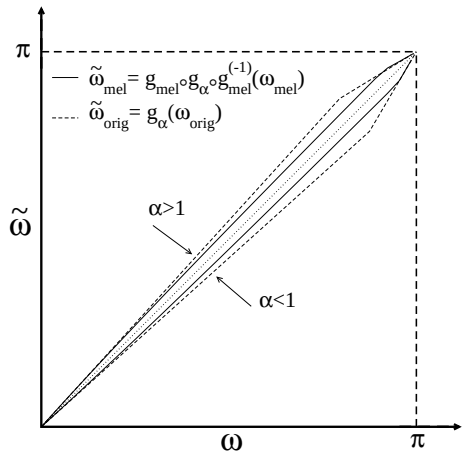

Figure 7.11: Effective warping function $g_{\text{eff}}^{(-1)} = g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}$ for combined Mel and VTN warping as function of the Mel frequency ω_{mel} (straight) in comparison to the warping function g_α for plain CC as function of the original frequency ω_{orig} (dashed).

$\tilde{\omega}_{\text{mel}} = \left(g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}\right)(\omega_{\text{mel}})$ (straight line in Fig. 7.11) as function of the Mel frequency with the warping function g_α for plain CC (dashed line) as function of the original frequency reveals that $g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}$ is much closer to identity than g_α. Thus the transformation matrix is expected to be even more diagonally dominant than those obtained for plain Cepstral coefficients. Transformation matrices for MFCC using piece-wise linear warping with warping factors ($\alpha = 0.9$ and $\alpha = 1.1$) are shown

in Fig. 7.12 and 7.13. These matrices were calculated by solving Eq. (7.39) numerically without approximations. The figures show that the transformation matrices for MFCC are indeed dominated more by the diagonal elements than the matrices for plain CC (Fig. 7.5 and 7.6). Thus the transformation matrix for MFCC may be approximated by a *tri*diagonal matrix rather than a *quin*diagonal matrix (cf. Section 7.4). A first experimental evaluation is given in Section 9.3.

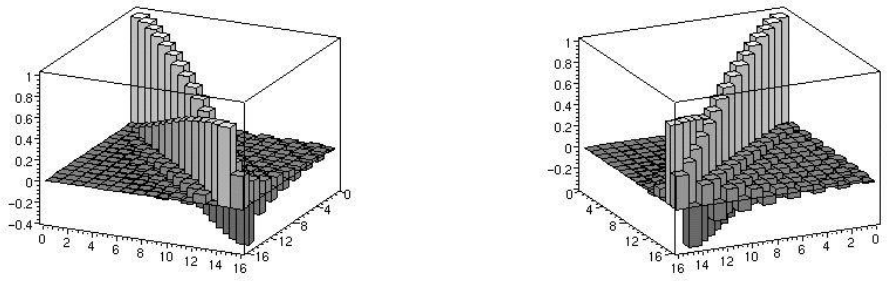

Figure 7.12: Matrix for piece-wise linear warping function, $\alpha = 0.9$, Mel frequency scale.

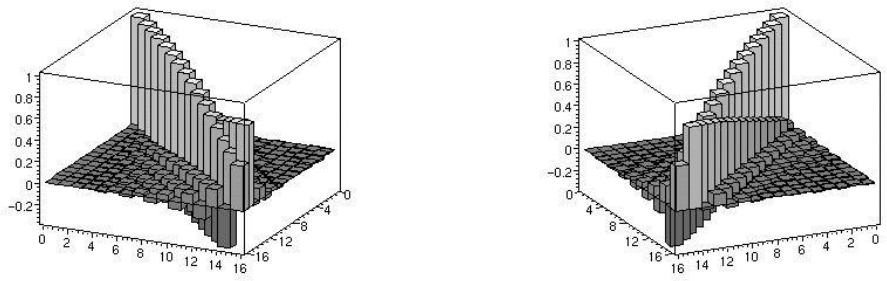

Figure 7.13: Matrix for piece-wise linear warping function, $\alpha = 1.1$, Mel frequency scale.

7.6 Examples of Spectra Warped by Linear Transformation

In the traditional or the integrated signal analysis approach (c.f. Section 7.2), the smoothing has to be applied by omitting higher *warped* Cepstral coefficients because VTN is applied before the Cepstrum transformation or integrated into the Cepstrum transformation. When using the transformation matrix calculated in this Chapter, the smoothing can be applied at two different stages: to the *warped* Cepstral coefficients (Scheme a) in Fig. 7.14) or to the *unwarped* Cepstral coefficients (Scheme b)

in Fig. 7.14). Smoothing by omitting higher unwarped Cepstral coefficients (Scheme b)) results in a loss of frequency resolution but has the advantage of faster computation because less unwarped Cepstral coefficients have to be calculated. Additionally, the multiplication of the matrix $A_{nk}(\alpha)$ with the Cepstral vector is faster because $A_{nk}(\alpha)$ is of lower dimension. The differences between both smoothing schemes will be discussed in the following. A sample spectrum (Fig. 7.15, $\alpha = 1.0$) with $N = 512$

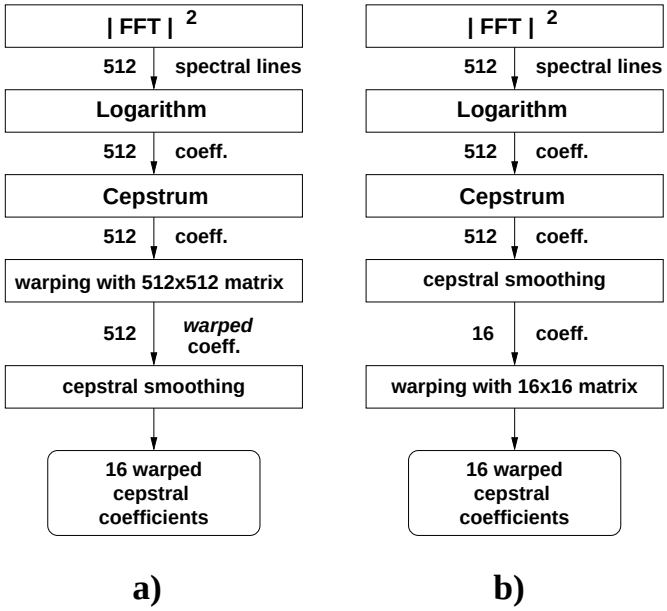

Figure 7.14: Schemes for different stages of Cepstral smoothing: a) smoothing after VTN warping, b) smoothing before VTN warping.

spectral lines was transformed into $K = 512$ Cepstral coefficients by a Cepstrum transformation:

$$c_k = \frac{4}{N} \sum_{n=0}^{N/2-1} \ln \left| X(\mathrm{e}^{i\frac{2\pi n}{N}}) \right|^2 \cos(\frac{2\pi n}{N}\, k) \; .$$

Then the Cepstral vector has been transformed using Eq. (7.9) into a piece-wise linear warped Cepstral vector of 512 coefficients for warping factors $\alpha = 0.8$ and $\alpha = 1.2$, respectively. Afterwards, the inverse DCT has been applied to the warped Cepstral vector in order to obtain a warped spectrum. This last transformation has been carried out for demonstration only; in practice the warped Cepstral vector is used for further processing. The resulting spectra are shown in Fig. 7.15. A comparison of the warped Cepstral coefficients obtained by the method presented here with those computed from the spectrum using the integrated approach described in Chapter 6 reveals no differences. As an additional example the effect of Cepstral smoothing is shown in Fig. 7.16. Again, the spectrum shown in Fig. 7.15 has been

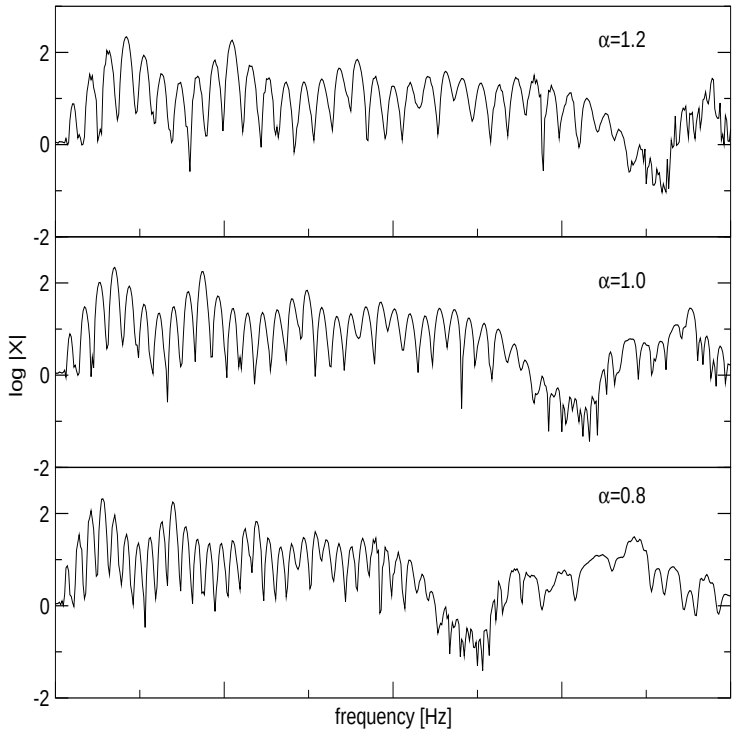

Figure 7.15: Example of warped spectra with warping factors $\alpha = 0.8$ and $\alpha = 1.2$.

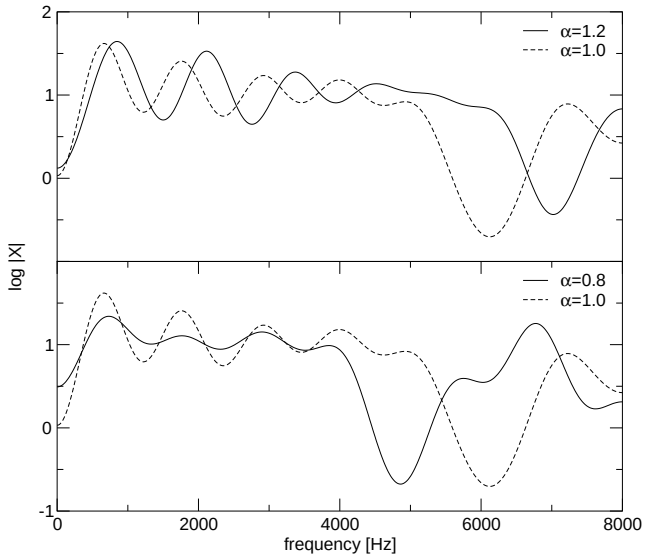

Figure 7.16: Example of a smoothed spectrum; the Cepstrum was warped with a 512×512 matrix and subsequently reduced to 16 coefficients (Scheme a).

transformed into 512 Cepstral coefficients and has now been smoothed by trans-

forming back with only the first 16 Cepstral coefficients ($\alpha = 1.0$ in Fig. 7.16). The warped spectra have been obtained by calculating 512 Cepstral coefficients, transforming them with Eq. (7.9) into 512 warped Cepstral coefficients, and subsequent smoothing by transforming back with only the first 16 warped Cepstral coefficients (Scheme a) in Fig. 7.14).)

As stated before, the warping obtained from the integrated approach of Chapter 6 can be reproduced only if Scheme a) is applied. If smoothing is applied first by calculating only the first 16 Cepstral coefficients and warping hereafter using a 16×16 matrix (Scheme b)), slightly different results are obtained. The difference between both methods is shown in Fig. 7.17. These differences arise mainly in the part of the spectrum where the slope of the warping function is larger than 1 (cf. Fig. 7.2), i.e. for $\omega > \omega_0$ for $\alpha = 0.8$ and $\omega < \omega_0$ for $\alpha = 1.2$ ($\omega_0 = \frac{7}{8}\pi \hat{=} 7000$Hz).

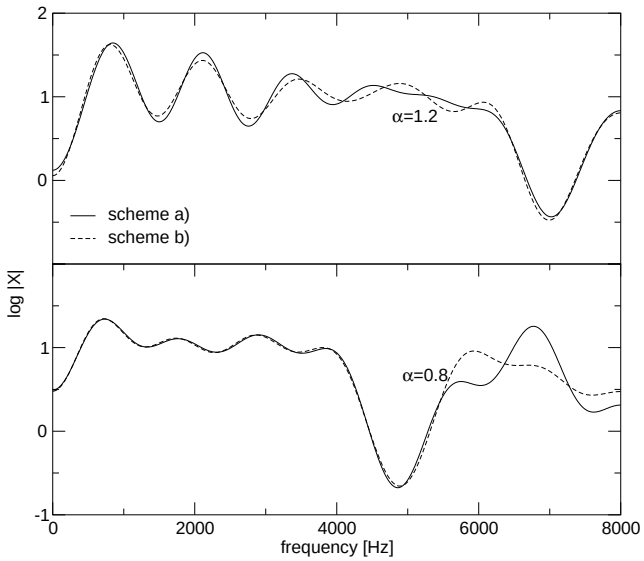

Figure 7.17: Effect of different order of warping and smoothing on the smoothed spectrum.

Recognition test results using Scheme a) and Scheme b) from Fig. 7.14 on the Verbmobil II corpus are given in Table 7.1. The experimental setup for the following experiments is as follows

- 512 Mel-warped Cepstral coefficients using the integrated approach as presented in Chapter 6
- 16 VTN warped Cepstral coefficients as presented in Section 7.5.2, augmented with first derivatives and second derivative of the energy
- short-term Cepstral mean normalization on a sliding window of 2 seconds length

Table 7.1: Recognition test results on Verbmobil II. Warping factor estimation according to Scheme a) and b) of Fig. 7.14.

MFCC comp. scheme	VTN	WER
Scheme a)	no	23.1%
	yes	21.1%
Scheme b)	no	23.2%
	yes	21.2%

- first order derivatives (using linear regression on five frames), second order derivative of energy
- LDA of three adjacent augmented VTN-MFCC vectors with reduction to 33 dimensions
- no variance and energy normalization
- 3500 decision-tree clustered across-word HMM states
- 3 state HMM topology

The acoustic model training was performed according to the RWTH standard training as described in [Molau 03, p.71][Sixtus 03, p.53]

Although the spectra obtained by both schemes differ slightly (c.f. Fig. 7.17), the recognition results differ only within the range of statistical significance.

7.7 Interdependence of VTN and MLLR

Adaptation and normalization are commonly viewed as different techniques to reduce the mismatch between training and testing conditions [Woodland 01]. In Chapter 4 it has been shown that adaptation and normalization can be put together in the same mathematical framework. Moreover they are identical in terms of Bayes' decision rule. As a consequence of VTN being a linear transformation of the Cepstral vector, a strong interdependence of two widely used normalization (VTN) and adaptation (MLLR) techniques can be derived if Gaussian emission probabilities are used.

Most of today's ASR systems make use of Hidden Markov Models (HMM) with Gaussian emission probability distributions

$$\mathcal{N}(x|\mu,\mathbf{\Sigma}) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}} \exp\left\{-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right\} .$$

with state dependent parameters μ and Σ. If the acoustic feature vector (essentially the Cepstral coefficients) x is normalized with the VTN matrix $\mathbf{A}$, the Gaussian distribution changes to

$$\begin{aligned} x \rightarrow y &= \mathbf{A}\, x : \\ \mathcal{N}(x|\mu, \mathbf{\Sigma}) &\rightarrow \mathcal{N}(y|\mu, \mathbf{\Sigma}) \\ &= \mathcal{N}(x|\boldsymbol{A}^{-1}\,\mu, \boldsymbol{A}^{-1}\,\mathbf{\Sigma}\,\boldsymbol{A}^{-1^T}) \\ &= \mathcal{N}(x|\hat{\mu}, \hat{\mathbf{\Sigma}}) \end{aligned}$$

with

$$\hat{\mu} = \boldsymbol{A}^{-1}\,\mu \qquad \text{and} \qquad \hat{\mathbf{\Sigma}} = \boldsymbol{A}^{-1}\,\mathbf{\Sigma}\,\boldsymbol{A}^{-1^T}\ . \tag{7.42}$$

Thus, a linear transformation of the observation vector x is equivalent to a linear transformation of the mean vector μ and an appropriate transformation of the covariance matrix $\mathbf{\Sigma}$. This topic has already be discussed in a more general view in Chapter 5.

The transformations in Eq. (7.42) describe a constrained MLLR (C-MLLR) [Digalakis & Rtischev$^+$ 95, Gales 98]. The attribute constrained refers to the use of the same matrix $\mathbf{A}$ for the transformation of the mean and variance, c.f. Section 5.2.

In [Uebel & Woodland 99], Uebel and Woodland have found empirically that improvements obtained by C-MLLR and VTN were not additive. As shown above, VTN may be viewed as a special case of C-MLLR adaptation with an restriction to only one adjustable parameter (the warping parameter) that determines the matrix elements. The experiments were based on a MF-PLP signal analysis. The difference between MFCC and MF-PLP is mainly caused by different types of smoothing, which is not expected to effect the equivalence of VTN and linear transformations. Hence, the experiments support the analytic result that VTN is a special case of C-MLLR.

7.8 Conclusions

Currently, the warping factor is estimated similarly in nearly all VTN approaches: The spectrum is warped with a discrete set of warping factors and a forced alignment of the utterance is performed using a given transcription (either the correct one or obtained by a preliminary recognition pass). The warping factor for which the alignment yields the best score is then chosen. This procedure is quite time consuming since a signal analysis and a time alignment for each warping factor in the set is required. Although there are some approaches to speed up the warping factor estimation [Welling & Ney$^+$ 02], calculating the warping factor analytically from the Cepstral vectors would be preferable. As the transformation matrix can

now be calculated analytically, the warping factor could be directly calculated rather than being estimated by a grid search. Unfortunately, the resulting objective function for an HMM with Gaussian emission probability with variance $\mathbf{\Sigma}_s$ and mean μ_s and posterior probability $\gamma_s(t)$ for being in state s at time t

$$\begin{aligned}\alpha^{opt} = \underset{\alpha}{\operatorname{argmax}} \sum_{t=1}^{T} \sum_{s=1}^{S} \Bigg\{ \gamma_s(t) \Big(& \log |\det \boldsymbol{A}(\alpha)| \\ & - \frac{1}{2} (\boldsymbol{A}(\alpha)\, x_t - \mu_s)^T \mathbf{\Sigma}_s^{-1} (\boldsymbol{A}(\alpha)\, x_t - \mu_s) \Big) \Bigg\}\end{aligned} \tag{7.43}$$

can hardly be solved explicitly, especially for MFCC (cf. Eq. (7.39)).

Using the standard grid search approach for estimating the warping factor, the approximation of the transformation matrix derived in Sections 7.4 and 7.5 may be of less importance. In that case, the computing the approximated matrix has similar costs as computing the complete matrix A_{nk}. But the approximation may be useful in another way: as shown in Section 7.7, VTN can be considered as a special case of C-MLLR. Thus, the approximation motivated in Sections 7.4 and 7.5 can be used for the case of a feature space transform as described in [Gales 98]. In that case, a tridiagonal matrix could result in a similar performance with much less parameters.

A first exploitation of the theoretical results presented in this Chapter is given in Section 9.3. Experimental studies show that a band diagonal restriction of the matrix used for MLLR adaptation clearly outperforms the full transform for small amounts of adaptation data and is only slightly inferior to the full transform for large amounts of adaptation data.

7.9 Summary

In this Chapter it has been shown that vocal tract normalization can be expressed as a linear transformation of the Cepstral vector for *arbitrary* invertible warping functions. For the case of piece-wise linear, quadratic and bilinear warping an analytic solution for the transformation matrix has been derived exemplary. Using special properties of typical warping functions it has been shown that the transformation matrix can be approximated by a quindiagonal matrix for plain Cepstral coefficients or a tridiagonal matrix for MFCC, respectively. Computing the transformation matrix for VTN allows a proper normalization of the probability distribution with the Jacobian determinant of the transformation. Finally it has been illustrated that VTN amounts to a special case of MLLR. This explains previous experimental results that improvements obtained by VTN and subsequent C-MLLR were not additive, which had not been understood so far.

Chapter 8

Effect of the Jacobian Determinant on Vocal Tract Normalization

In Section 4.3 was discussed that the Jacobian determinant needs to be taken into account for vocal tract normalization (VTN). In virtually all approaches, the Jacobian determinant is assumed to be flat as function of the warping parameter and hence neglected. In [McDonough 00] it was shown that the Jacobian determinant plays an important role if an all-pass transform is used as warping function. With the approach presented in Chapter 7 it is for the first time possible to study the effect of the Jacobian determinant on VTN in general. Although VTN works quite well without taking the Jacobian determinant into account, a systematic study of the effects has not been done yet. The correct treatment should result in an improved warping factor estimation and thus improved recognition accuracy.

8.1 Jacobian Determinant for Vocal Tract Normalization

To estimate the unknown warping factor α, the procedure is as follows [Welling & Ney$^+$ 02]: For each speaker r, labeled training data $(x_{r\,1}^T, w_{r\,1}^N)$ are given, where $x_{r\,1}^T$ denotes the sequence of acoustic data and $w_{r\,1}^N$ the sequence of spoken words. In recognition, a preliminary hypothesis of the unknown word sequence $w_{r\,1}^N$ can be obtained by a first recognition pass. Typically, a maximum likelihood estimation is applied to obtain the unknown warping factor α of a feature transformation function g_α [Sankar & Lee 96]

$$\hat{\alpha}_r = \underset{\alpha}{\operatorname{argmax}} \left\{ p(g_\alpha(x_{1\,r}^T)|w_{1\,r}^N) \cdot \left| \frac{dg_\alpha(x_{1\,r}^T)}{dx_{1\,r}^T} \right| \right\} \ . \tag{8.1}$$

The last term denotes the Jacobian determinant of the transformation. It can be omitted if no direct comparison of probability values is carried out with differently "normalized" distributions. In a typical VTN approach the warping factor α is chosen by comparing the scores obtained by a forced alignment of the same utterance

warped with a set of discrete warping factors. Hence the Jacobian determinant needs to be taken into account [Sankar & Lee 96]. Often the Jacobian determinant is assumed to be flat as function of α. Accordingly it is approximated to be independent of α and thus neglected.

In VTN the speaker normalization is usually not performed as a transformation of the acoustic vectors but by warping the power spectrum during signal analysis instead, and the Jacobian determinant can hardly be calculated. In virtually all experimental studies the second factor in Eq. (8.1), the Jacobian determinant, is neglected. Whether this is a good approximation or not will depend very much on how much the Jacobian determinant depends on α. Therefore it is good to study the second term as function of α. By expressing VTN as a matrix transformation of the acoustic vectors ($x \rightarrow \mathbf{A}\,x$) it is for the first time possible to study the Jacobian determinant $|\det \mathbf{A}|$ for all kinds of transformation functions. Taking the negative logarithm of Eq. (8.1), the Jacobian determinant results in an additive term to the acoustic score $-\log p(g_\alpha(x_{1\,r}^T)|w_{1\,r}^N)$, thus the maximum likelihood estimation of the warping factor α reads

$$\hat{\alpha}_r = \underset{\alpha}{\operatorname{argmin}} \left\{ -\log p(g_\alpha(x_{1\,r}^T)|w_{1\,r}^N) - \log|\det \mathbf{A}| \right\} \ . \tag{8.2}$$

8.2 Effect on Warping Factor Estimation and Experimental Results

The goal of the following investigation is to study if the conventional VTN implementation can be improved by incorporating the Jacobian determinant without changing the conventional signal analysis. As shown in Section 7.4, the shape of the transformation matrices is very similar for different functional forms of the warping function g_α because of common characteristics of the warping functions. Additionally, in [Molau & Kanthak$^+$ 00] a study of different warping functions revealed little impact on the recognition accuracy. As the piece-wise linear warping function is the most popular warping function, the investigation is focused on that functional form. An investigation of other warping functions is straightforward with the analytical solutions given in Chapter 7.

In Section 7.6 two different schemes for computing warped Cepstral coefficients have been discussed. For convenience, Fig. 7.14 is shown again in this Chapter as Fig. 8.1. Scheme a) yields the same Cepstral coefficients as the signal analysis presented in Chapter 6. In that Scheme, the resulting Cepstral coefficients are obtained by transforming the 512-dimensional vector of Cepstral coefficients with a 16×512 matrix, i.e. includes a dimensionality reduction. For that case, the Jacobian determinant is difficult to calculate. On the other hand, Scheme b) yields very similar Cepstral coefficients with a 16×16 dimensional transformation matrix and thus it is assumed that Jacobian determinant for Scheme a) can be approximated by the

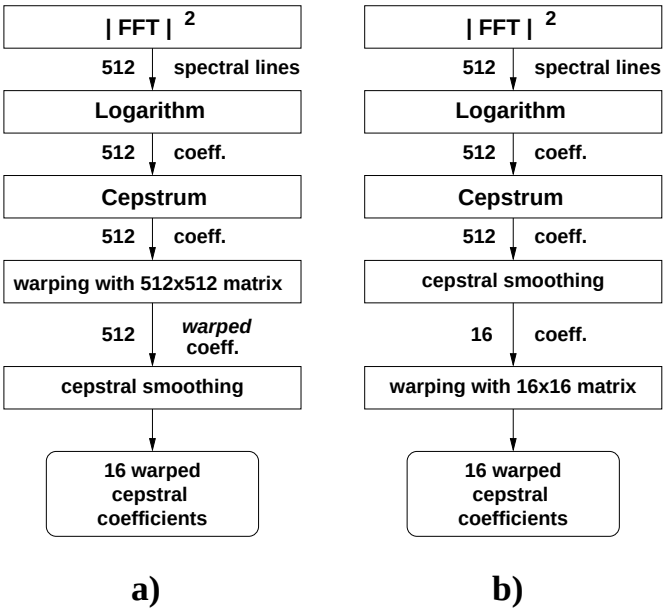

Figure 8.1: Schemes for different stages of Cepstral smoothing: a) smoothing after VTN warping, b) smoothing before VTN warping

Jacobian determinant for Scheme b). For the latter case, the Jacobian determinant is simply the determinant of the 16×16 dimensional transformation matrix. Therefore in this Section the term Jacobian determinant refers *always* to the determinant of the 16×16 matrix from Scheme b). In other words, the Jacobian determinant which has been calculated for a 16×16 matrix has been also used to modify the warping factors obtained for Scheme a), i.e. by using a 16×512 matrix. It should be clearly stated that using the Jacobian determinant from Scheme a) for a signal analysis according to Scheme b) is an approximation, which has been introduced for two reasons: firstly, the Jacobian determinant for Scheme a) can hardly be calculated. Secondly, the goal is to account for the Jacobian determinant while altering the signal analysis as little as possible. Therefore the following study analyzes the effect of using the Jacobian determinant from Scheme b), i.e. a 16×16 transformation matrix, when using a signal analysis according to Scheme b), i.e. a 16×512 transformation matrix.

The experimental setup for the following experiments is as follows

- 512 Mel-warped Cepstral coefficients as presented in Chapter 6
- 16 VTN warped Cepstral coefficients as presented in Section 7.5.2, augmented with first derivatives and second derivative of the energy
- short-term Cepstral mean normalization on a sliding window of 2 seconds length

- first order derivatives (using linear regression on five frames), second order derivative of energy
- LDA of three adjacent augmented VTN-MFCC vectors with reduction to 33 dimensions
- no variance and energy normalization
- 3500 decision-tree clustered across-word HMM states
- 3 state HMM topology
- 10000 word lexicon

The acoustic model training was performed according to the RWTH standard training as described in [Molau 03, p.71][Sixtus 03, p.53].

Fig. 8.2 shows the Jacobian determinant which has been computed numerically as function of the warping factor α for a piece-wise linear warping function using the 16×16 matrix from Scheme b). The function $-\log|\det \mathbf{A}|$ penalizes warping factors with larger deviations from $\alpha = 1$ and thus keeps the transformation closer to identity. If Mel frequency warping is additionally applied (dashed line in Fig. 8.2), the dependence of the Jacobian determinant on α is much weaker because the effective warping function $g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}$ is much closer to identity than g_α for plain Cepstral coefficients (i.e. without Mel frequency warping) (cf. Fig. 7.11).

Fig. 8.3 shows a typical distribution of the log-likelihood $-\log p(g_\alpha(x_{r1}^T)|w_{r1}^N)$ (without the Jacobian determinant) for a female and a male speaker from the Verbmobil II corpus together with the Jacobian determinant $-\log|\det \mathbf{A}(\alpha)|$, all as function of the warping parameter α. Looking at this plot, it seems that the Jacobian determinant might play an important role for VTN as it is in the same order of magnitude as the log-likelihood itself. Thus, a correct consideration of the Jacobian determinant will result in major deviations in the estimation of the warping factors. The effect of the Jacobian determinant on the estimation of the warping factors is depicted in Fig. 8.4. Shown are the histograms of warping factors obtained using Scheme a) with and without taking the Jacobian determinant into account. It can be clearly seen that the histogram narrows towards the value $\alpha = 1$ (the identity transformation) when taking the Jacobian determinant into account. Although the consideration of the Jacobian determinant has a noticeable effect on the warping factor estimation, the effect on the word error rate is surprisingly small. Recognition results on the Verbmobil II corpus are given in Table 8.1. A slight degradation in the recognition accuracy is observed which is hardly beyond the significance limit. To care for possible systematic errors introduced by using the Jacobian determinant from Scheme b) for Scheme a), the value of the Jacobian determinant has been scaled down by a certain factor resulting in a smaller change of the warping factors. The

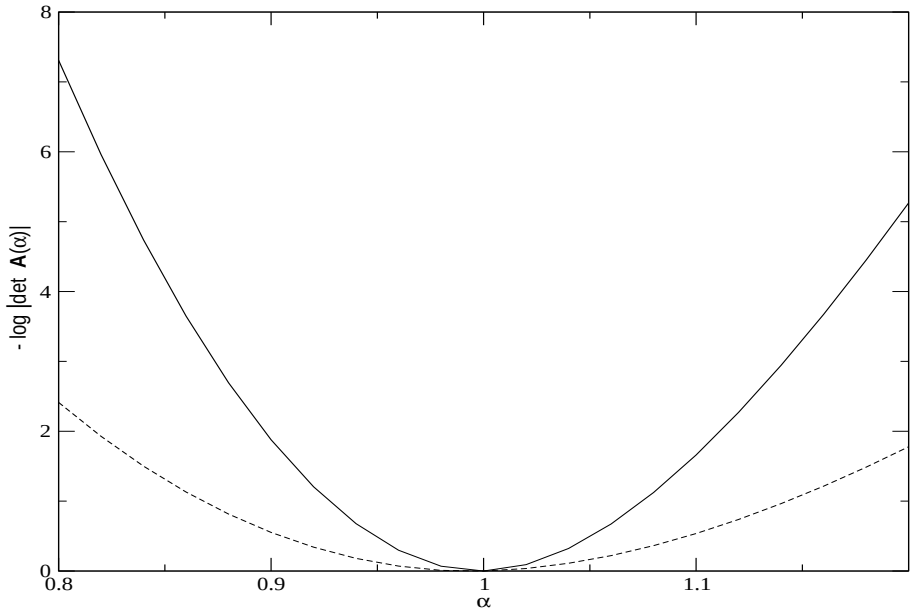

Figure 8.2: Plot of $-\log|\det \mathbf{A}|$ (Jacobian determinant) for piece-wise linear warping of 16 Cepstral coefficients as function of α. Straight line: original frequency scale, dashed line: Mel frequency scale

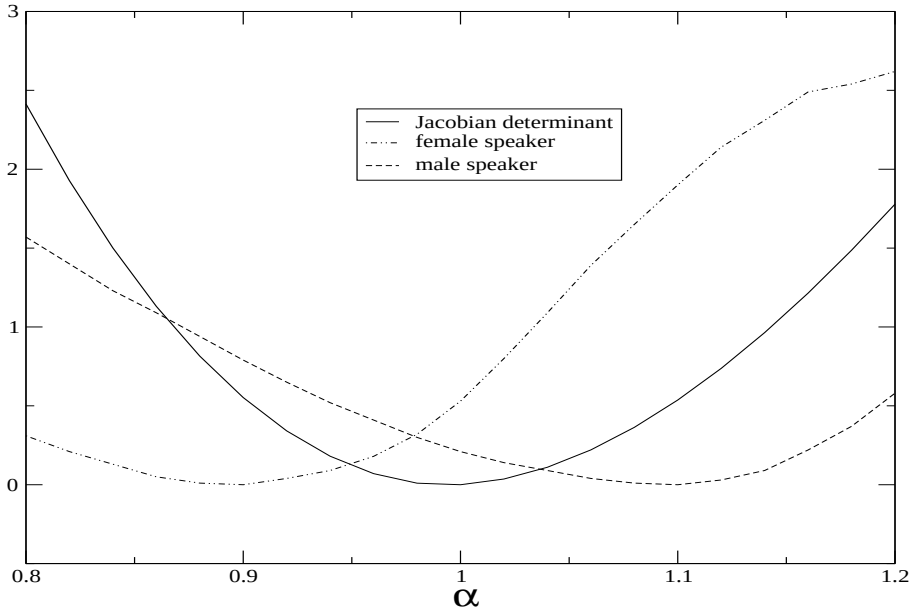

Figure 8.3: Normalized acoustic score (negative log-likelihood) of two test sentences from the Verbmobil II corpus as function of the warping factor α in comparison to the contribution of the Jacobian determinant $-\log|\det \mathbf{A}(\alpha)|$

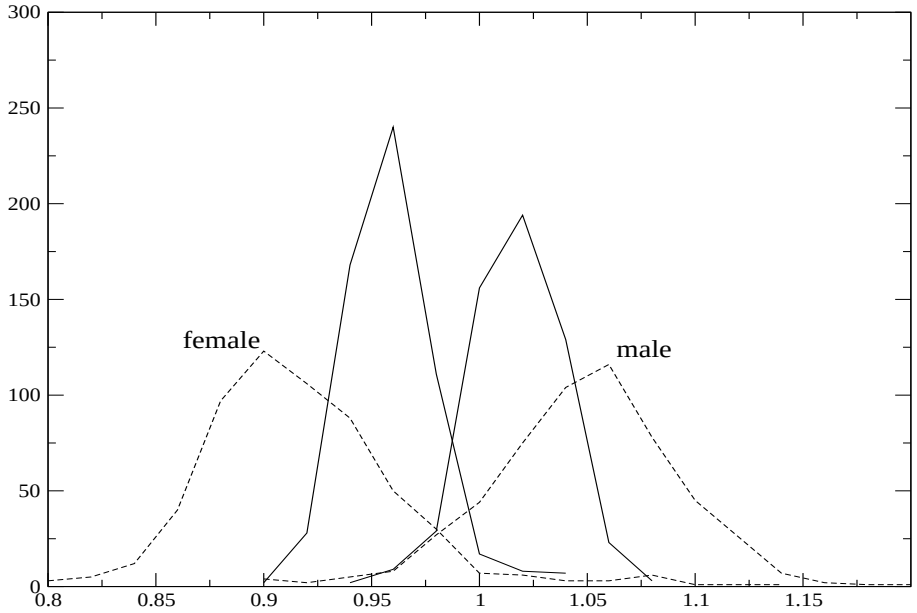

Figure 8.4: Histogram of estimated warping factors for the test speakers of the Verbmobil II corpus without (dashed line) and with (straight line) taking the Jacobian determinant into account. The warping factors were estimated using Scheme a) of Fig. 8.1.

Table 8.1: Recognition test results on Verbmobil II. Warping factor estimation according to Scheme a) of Fig. 8.1.

VTN	Jacobian determinant	scaling factor	WER
no			23.1%
yes	no	0.0	21.1%
	yes	1.0	21.3%
		0.5	20.8%
		0.25	21.0%

recognition accuracy is improved slightly, but again hardly beyond the significance limit.

As stated before, using the Jacobian determinant from Scheme b) of Fig. 8.1 for the signal analysis according to Scheme a) is an approximation. In order to study if the disappointing results shown in Table 8.1 are due to this approximation, a second set of experiments was carried out using the signal analysis from Scheme b) including the warping factor estimation, i.e. the complete signal analysis has been carried out according to Scheme b) and thus no approximation concerning the Jacobian determinant was made.

The histogram of the obtained warping factors is shown in Fig. 8.5. By comparing

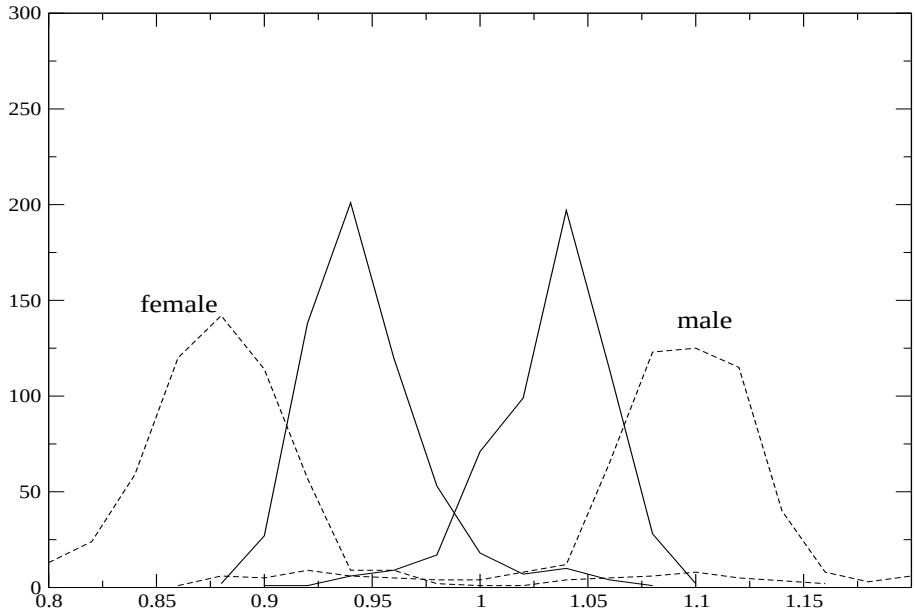

Figure 8.5: Histogram of estimated warping factors for the test speakers of the Verbmobil II corpus without (dashed line) and with (straight line) taking the Jacobian determinant into account. The warping factors were estimated using Scheme b) of Fig. 8.1.

this histograms with those depicted in Fig. 8.4, it can be seen that the latter are closer to the value $\alpha = 1$. This suggests that for the warping factors of Scheme b) a correction towards $\alpha = 1$ maybe helpful, indeed. Recognition results using the same experimental setup as before (except from the signal analysis) are given in Table 8.2. As expected, the baseline results are very similar. Incorporating the Jacobian determinant now results in a minor improvement (without the need of a scaling factor) but the effect is again very small.

Table 8.2: Recognition test results on Verbmobil II. Warping factor estimation according to Scheme b) of Fig. 8.1.

VTN	Jacobian determinant	WER
no		23.2%
yes	no	21.2%
	yes	20.9%

Surprisingly, the correct treatment of the Jacobian determinant has only little influence on the recognition accuracy despite the significant effect on the warping

factor histograms. A study of individual speakers as well as individual warping factors gave no further insight. In previous studies a correlation of the shape of the warping factor histograms with the recognition accuracy has been observed [Molau 02]. Thus, further research is needed on that topic.

In summary, it can be stated that the assumption of the small influence of the Jacobian determinant on the recognition accuracy is valid, at least for the Verbmobil II corpus, and that the approximation of dropping the Jacobian determinant is justified. As the warping factor histograms and functional forms of the log-likelihood are similar for other corpora, the same result can be expected.

8.3 Discussion

The warping factor α is usually determined by warping the speech signal with a discrete set of warping factors and choosing the most likely one according to Eq. (8.1), neglecting the Jacobian determinant. With the limitation of possible warping factors to a set of discrete values close to the identity transformation, the span of possible transformations is reduced. Thus the errors caused by improper normalization of the probability distribution by neglecting the Jacobian determinant are small and do not cause VTN to fail.

When not restricting the span of possible transformations, for instance in Maximum Likelihood Linear Transforms (MLLT) [Gales 98], the normalization of the acoustic feature vectors will fail without proper normalization of the probability distributions. Experiments have shown that unrestricted linear transformation with neglecting the Jacobian determinant cause the automatic speech recognition system to spuriously recognize silence only. A similar result has been reported in [Cox 00]. In the following an empirical explanation of these results will be given. A transformation of the acoustic feature vectors using a matrix $\boldsymbol{A}$ without any restrictions to the values of the matrix elements while neglecting the Jacobian determinant results in a density function

$$d_A(x|\mu,\Sigma) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\mathbf{A}x-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{A}x-\mu)\right\} ,$$

which is not normalized. If the transformation matrix $\mathbf{A}$ is estimated using a maximum (log)likelihood criterion, the following will most probably happen: Among the possible means μ of the ASR system, one or more means are assigned to the *silence* model, which contains usually low spectral energy and thus small Cepstral coefficients. The transformation will now map the acoustic feature vectors onto the silence model by simply shrinking the values of the Cepstral coefficients[1]. As a sim-

[1] If the silence model is treated separately, the estimated transformation will map the feature vectors to the mean of a consonant model with low spectral energy

plified example, assume that the mean vector of the silence model consists of zeroes only. If the transformation matrix is identical to $\mathbf{0}$, all transformed feature vectors match perfectly with the silence model independent of their actual values. However, if the density is properly normalized by taking the Jacobian determinant into account, this results in an additional term $\log|\det \mathbf{A}|$ in the log-likelihood function:

$$\log \mathcal{L}(x_1^T; d_A(x|\mu,\Sigma)) = \sum_{t=1}^{T} \log d_A(x_t|\mu,\Sigma) + T \cdot \log|\det \mathbf{A}|$$

This second term $\log|\det \mathbf{A}|$ plays the role of a penalty for the transformation onto the silence model because it penalizes transformation to small Cepstral values. In the example above, this would result in an infinite penalty which balances the perfect match of the first term.

The reason why VTN (in contrast to MLLT) does not fail without proper normalization lies in the limitation of the warping factors and the specification of the warping function. The resulting transformation has not enough degrees of freedom and thus the matrix is kept close to the identity matrix.

8.4 Summary

The representation of vocal tract normalization as a linear transformation of the Cepstral coefficients, as shown in Chapter 7, allows for the first time a detailed analysis of the influence of the Jacobian determinant on the VTN warping factor estimation. In this Chapter this has been investigated exemplary for a piece-wise linear warping function, which is used in many VTN approaches. In nearly all previous VTN implementations the Jacobian determinant has been neglected. It turned out that the Jacobian determinant has a significant influence on the estimation of the warping factors, which resulted in a major difference between the particular the warping factor histograms. Surprisingly, the recognition performance turned out to be almost unaffected, despite of the major changes in the values of the warping factors. In general, the approximation of neglecting the Jacobian determinant has been shown to be justified. However, further research is needed to understand the small impact of the Jacobian determinant on the recognition performance given the large impact on the warping factor histograms.

Chapter 9

Maximum Likelihood Linear Regression

9.1 Introduction

Maximum likelihood linear regression (MLLR) [Leggetter & Woodland 95a] belongs to the transformation family of adaptive acoustic modeling (cf. Section 4.2.2). The basics of MLLR adaptation are given in Section 5.2.1 and are repeated in this Section for completeness.

The MLLR approach uses an affine transformation to adjust the mean vectors μ_s of the emission probability distributions to a new speaker or environmental condition:

$$\hat{\mu}_{s,r} = \boldsymbol{A}_{s,r}\,\mu_s + b_{s,r} \tag{9.1}$$

where s denotes the HMM state and r the speaker (or condition). For notational convenience, Eq. (9.1) is usually rewritten in the form

$$\hat{\mu}_{s,r} = \boldsymbol{W}_{s,r}\,\xi_s \tag{9.2}$$

where ξ_s denotes the extended mean vector

$$\xi_s = [1\ \mu_s^\top]^\top \tag{9.3}$$

and $\boldsymbol{W}_{s,r}$ is the $n \times (n+1)$-matrix $[b_{s,r}\,\boldsymbol{A}_{s,r}]$. The adaptation matrix $\boldsymbol{W}_{s,r}$ is estimated by maximum likelihood, given adaptation data $({x_r}_1^T, {w_r}_1^N)$:

$$\boldsymbol{W}_{s,r}^{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{W}_{s,r}} p({x_r}_1^T | {w_r}_1^N, \theta, \boldsymbol{W}_{s,r}) \tag{9.4}$$

The maximization of Eq. (9.4) is carried out using the expectation-maximization (EM) algorithm [Dempster & Laird$^+$ 77] and results in the following optimization problem:

$$\boldsymbol{W}_{s,r}^{\mathrm{ML}} = \operatorname*{argmin}_{\boldsymbol{W}_{s,r}} \left\{ \sum_{t=1}^{T} \gamma_s(t)(x_{rt} - \boldsymbol{W}_{s,r}\,\xi_s)^\top \boldsymbol{\Sigma}_s^{-1}(x_{rt} - \boldsymbol{W}_{s,r}\,\xi_s) \right\} . \tag{9.5}$$

For diagonal covariance matrices, which are used in the RWTH system, taking the derivative w.r.t. $\boldsymbol{W}_{s,r}$ and equating to zero yields a row-wise solution for $\boldsymbol{W}_{s,r}^{\mathrm{ML}}$ [Leggetter & Woodland 95a]:

$$W_{s,r}^{(i)} = Z_s^{(i)} \boldsymbol{G}_s^{(i)^{-1}} \tag{9.6}$$

with

$$\boldsymbol{G}_s^{(i)} = \frac{1}{\sigma_{s\,i}^2} \xi_s \xi_s^\top \sum_{t=1}^{T} \gamma_s(t) \tag{9.7}$$

$$Z_s^{(i)} = \sum_{t=1}^{T} \gamma_s(t) \frac{1}{\sigma_{s\,i}^2} x_{rt} \xi_s^\top \tag{9.8}$$

where $W_{s,r}^{(i)}$ and $Z_s^{(i)}$ denote the i-th row-vector of $\boldsymbol{W}_{s,r}^{\mathrm{ML}}$ and $\boldsymbol{Z}_s$, respectively, and $\sigma_{s\,i}^2$ is the i-th diagonal element of $\boldsymbol{\Sigma}_s$. A solution for the full covariance case is given in [Gales & Woodland 96]. The main advantage of MLLR is that several HMM states may share the same transformation matrix and thus even HMM states which are unseen or have a small number of observations in the adaptation data can be adapted. Usually, the adaptation matrices are tied over several HMM states thus defining regression classes $c = 1, \ldots, C$. Thus, Eqs. (9.7)–(9.8) read

$$\boldsymbol{G}_c^{(i)} = \sum_{m=1}^{M} \frac{1}{\sigma_{s_m\,i}^2} \xi_{s_m} \xi_{s_m}^\top \sum_{t=1}^{T} \gamma_{s_m}(t) \tag{9.9}$$

$$Z_c^{(i)} = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{s_m}(t) \frac{1}{\sigma_{s_m\,i}^2} x_{rt} \xi_{s_m}^\top \tag{9.10}$$

where the sum over m runs over all states $\mathcal{S}_c = \{s_1, \ldots, s_m, \ldots, s_M\}$ which belong to class c.

The RWTH system applies the Viterbi approximation, and the covariances $\boldsymbol{\Sigma}_s$ are globally tied over all HMM states. Hence, Eqs. (9.5), (9.9)–(9.10) simplify to

$$\boldsymbol{W}_{c,r}^{\mathrm{ML}} = \boldsymbol{Z}_c \boldsymbol{G}_c^{-1} \tag{9.11}$$

$$\boldsymbol{G}_c = \sum_{\substack{t=1 \\ s_t \in c}}^{T} \xi_{s_t} \xi_{s_t}^\top \tag{9.12}$$

$$\boldsymbol{Z}_c = \sum_{\substack{t=1 \\ s_t \in c}}^{T} x_{rt} \xi_{s_t}^\top \tag{9.13}$$

9.2 Modeling of Regression Classes

One of the main advantages of adaptation schemes belonging to the transformation family (c.f. Section 4.2.2) is the use of regression classes. The number of transformations used to adapt the speech recognition system to a new speaker and/or environment is chosen dependent on the available adaptation data, for instance a global transformation is used for the case of limited adaptation data and one transformation for each phoneme is applied if many adaptation data is available. This Section deals with the definition of regression classes and presents some refined regression class definition, especially for the case of limited adaptation data.

9.2.1 Refined Bias Modeling

In the conventional approach [Leggetter & Woodland 95a], an equal number of matrices and bias vectors is used for adaptation. It has been shown, however, that a more detailed approach can lead to better adaptation results [Digalakis & Berkowitz$^+$ 99]. In the following a study on the effect of using different numbers of classes for the MLLR matrix $\boldsymbol{A}$ and the bias vector b is presented:

$$\hat{\mu}_s = \boldsymbol{A}_{c'}\mu_s + b_c \ ,$$

where the adaptation class c' is a function of the more detailed classes c

$$c' = c'(c) \ ,$$

and the refined classes c themselves are functions of the HMM states $c = c(s)$ (i.e. several HMM states share the same adaptation class). The estimation formula remain almost the same, only the estimation of the bias vector b (which is the first column in the extended matrix $\boldsymbol{W}$) has to be refined:

$$b_c = \frac{1}{T}\sum_{\substack{t=1 \\ s_t \in c}}^{T} \left(x_t - \boldsymbol{A}_{c'(c)}\mu_{s_t}\right) \tag{9.14}$$

The experimental setup is as follows:

- 16 Mel-frequency Cepstral coefficients
- short-term Cepstral mean and variance normalization on a sliding window of 2 seconds length
- first order derivatives (using linear regression on five frames), second order derivative of energy
- LDA of three adjacent augmented MFCC vectors with reduction to 33 dimensions

- 2500 decision-tree clustered, within-word HMM states
- 243k Gaussian mixture densities
- 3 state HMM topology
- 10000 word lexicon
- two-pass unsupervised adaptation

Several modelings of adaptation classes were tested, ranging from one global transformation for all time frames up to three matrices/bias vectors for speech frames and one matrix/bias vector for silence/noise. Additionally, an adaptation class modeling using one matrix for speech and silence each and one bias vector for each phoneme plus silence was tested. The experimental results are summarized in Table 9.1. It can be clearly seen that using only few full matrices but more bias vectors outperforms an equal number of matrices and biases. With an equal number of matrices and bias vectors, the optimum was reached with 3 matrices and 3 bias vectors (WER = 23.2%) and could not be further improved by increasing the number of parameters. The use of more classes for the biases gave an additional gain of 2.6% rel. (WER = 22.6%) with approximately the same number of adaptation parameters. Using diagonal MLLR matrices could not improve the results, even when using many adaptation classes, in accordance with the results presented in [Leggetter & Woodland 95a]. Block matrices, which are widely used and are often superior to full matrices, are not meaningful when using an LDA transformation since there is no specific order of the Cepstral coefficients in the acoustic feature vector any more after the LDA transformation.

Table 9.1: Recognition test results on the Verbmobil II corpus for different modelings of adaptation classes.

Number of classes for		No. of adpt.	WER [%]
matrices W	bias vectors b	parameters	
-	-	-	24.6
1	1	1122	23.6
2	2	2244	23.4
3	3	3366	23.2
4	4	4488	23.5
2	50	3828	22.6

9.2.2 Dynamic Selection of Regression Classes

Several methods have been proposed to define the regression classes c of HMM states that share the same transformation matrix [Leggetter & Woodland 95b, Gales 96, Haeb-Umbach 01]:

- expert knowledge: the regression classes are divided into broad phonetic classes like nasals, fricatives, etc. defined by an expert
- distance in acoustic space: the mean vectors are clustered using bottom-up or top-down clustering based on their distance in the acoustic space.

The RWTH systems uses a combination of both methods: As the triphones are clustered using a decision tree based on phonetic questions anyway, the same tree is used for the MLLR adaptation. The number of leafs is usually too large for MLLR adaptation as most of the leafs will seldomly be observed in the adaptation data, the tree is cut at a certain level, usually in the order of magnitude of 100 leafs for a tree with a total of $5000 - 8000$ leaves. The tree cut to N leafs is carried out by storing the order of the splits during estimation of the tree in the training phase and using the first $N - 1$ splits only.

The regression classes are dynamically obtained using this pruned decision tree. The parameter to refine the regression classes is the number of observations for this regression class, which is subject to empirical optimization. A simple example of such a tree is given in Fig. 9.1. Using this tree approach, a single observation can

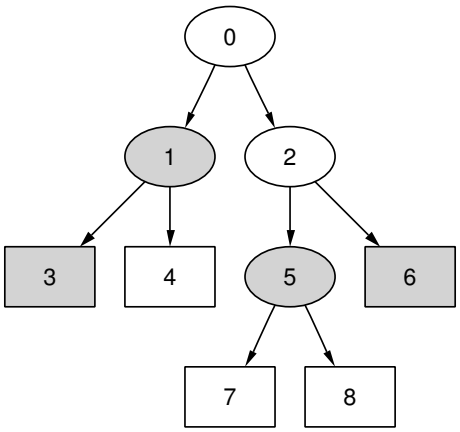

Figure 9.1: Example of MLLR regression class tree.

account for several regression classes. The statistics according to Eqs. (9.12)–(9.13) are estimated for the leaves of the tree during the enrollment of the adaptation data. Afterwards, the statistics are distributed across the complete tree, for instance

$$\boldsymbol{G}_1 = \boldsymbol{G}_3 + \boldsymbol{G}_4 \tag{9.15}$$

according to Fig. 9.1. The gray nodes in Fig. 9.1 denote that the associated regression class has enough observations, i.e. the number of observations seen in the adaptation data is larger than a certain threshold. Thus, HMM states which belong to class 3 are adapted with the MLLR matrix $\boldsymbol{W}_3$, estimated on data from class 3 only, and HMM states of class 4 are adapted using $\boldsymbol{W}_1$, estimated on adaptation data from both class 4 and class 3. The silence model is always adapted using a separate adaptation class. Experimental results of this approach on the Verbmobil II database are given in Table 9.2 for different threshold for the regression classes. The experimental setup is as follows:

- 16 Mel-frequency Cepstral coefficients
- short-term Cepstral mean and variance normalization on a sliding window of 2 seconds length
- first order derivatives (using linear regression on five frames), second order derivative of energy
- LDA of three adjacent augmented MFCC vectors with reduction to 33 dimensions
- 3500 decision-tree clustered, position-dependent across-word HMM states
- 360k Gaussian mixture densities
- 3 state HMM topology
- 10000 word lexicon
- two-pass unsupervised adaptation, CART tree used for definition of regression classes

Further details of the RWTH system can be found in [Sixtus 03]. The application of MLLR with one global adaptation matrix for speech and silence each gives a reduction in word error rate (WER) of 6.6% rel. The silence model is adapted because it is usually dependent on the environment and may additionally be speaker dependent (breath noises that are modeled by the silence model, for instance). Increasing the number of regression classes, the best performance is obtained at a threshold of 4000 observations per regression class, yielding a reduction in WER of 9.9% rel. These results are comparable with those reported in literature [Soltau & Schaaf$^+$ 01].

9.2.3 Semi-tied MLLR

The idea of semi-tied covariances (c.f. Section 5.3) can be transferred to MLLR adaptation. Based on the concept of Eigen-decomposition, in the semi-tied covariances approach the covariance matrices are decomposed into a state dependent diagonal matrix $\boldsymbol{\Sigma}_s^{\text{diag}}$ and a transformation $\boldsymbol{H}$ which is tied over all HMM states:

$$\boldsymbol{\Sigma}_s = \boldsymbol{H}\, \boldsymbol{\Sigma}_s^{\text{diag}}\, \boldsymbol{H}^\top \;, \tag{9.16}$$

Table 9.2: Recognition test results on the Verbmobil II corpus for different thresholds for the minimum number of observations per adaptation class using an adaptation class tree.

	min. obs.	Avg. no. matrices (regression classes)	WER %
baseline	–	0	22.15
MLLR	–	2 (fixed)	20.74
	500	45.7	20.84
	1000	35.2	20.73
	2000	24.9	20.25
	3000	19.2	20.09
	3500	16.9	20.14
	4000	**15.6**	**20.00**
	4500	14.3	20.05
	5000	12.4	20.26
	6000	10.6	20.35

where the matrix $\boldsymbol{H}$ is not necessarily orthogonal. Semi-tied covariances have been introduced to overcome the restriction of diagonal covariance matrices without the enormous increase in parameters to be estimated using a full covariance model.

A similar problem occurs with MLLR adaptation. When using adaptation classes, several matrices have to be estimated from possibly limited adaptation data. It was shown in [Leggetter & Woodland 95a] and confirmed by experiments with the RWTH system that a few full matrices are superior to many diagonal matrices. Thus, the off-diagonal elements of the MLLR transformation matrix play an important role for the adaptation performance. On the other hand, MLLR adaptation with only one global matrix gives already good improvements. The idea of semi-tied MLLR adaptation is now as follows: use a global transformation (i.e. tied over all regression classes) and a class dependent diagonal MLLR matrix. In other words, the model space is transformed into a new space where the assumption of a diagonal MLLR transformation matrix is more valid. As the MLLR matrix is not symmetric in general, a decomposition based on the Eigen-decomposition may not be used. Therefore, the MLLR matrix is decomposed similar to the *Singular Value Decomposition (SVD)*:

$$\boldsymbol{A}_c = \boldsymbol{U}\boldsymbol{\Lambda}_c\boldsymbol{V}^\top \tag{9.17}$$

where $\boldsymbol{\Lambda}_c$ is diagonal. Again, the transformation matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ do not have to be orthogonal. Using the Viterbi approximation and globally pooled covariances,

the following optimization problem has to be solved:

$$\underset{\{\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{V}\}}{\operatorname{argmin}} \sum_{t=1}^{T} \|x_t - \boldsymbol{A}_c \mu_{s_t}\|^2 = \underset{\{\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{V}\}}{\operatorname{argmin}} \left\{ \operatorname{tr}(\boldsymbol{A}_c{}^\top \boldsymbol{A}_c \, \boldsymbol{G}_c) - 2 \operatorname{tr}(\boldsymbol{A}_c{}^\top \boldsymbol{Z}_c) \right\} \tag{9.18}$$

with

$$\boldsymbol{A}_c = \boldsymbol{U} \, \boldsymbol{\Lambda}_c \, \boldsymbol{V}^\top \qquad \text{and} \qquad \boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_c \mid c = 1, \ldots, C\} \; .$$

and $\boldsymbol{G}_c$ and $\boldsymbol{Z}_c$ are defined similar to Eqs. (9.12) and (9.13):

$$\boldsymbol{G}_c = \sum_{\substack{t=1 \\ s_t \in c}}^{T} \mu_{s_t} \mu_{s_t}^\top \tag{9.19}$$

$$\boldsymbol{Z}_c = \sum_{\substack{t=1 \\ s_t \in c}}^{T} x_{rt} \mu_{s_t}^\top \tag{9.20}$$

The bias term b_c will be dropped for the moment to simplify the equations, because the estimation of that term is not changed by this approach. Taking the derivatives with respect to $\boldsymbol{\Lambda}_c$, $\boldsymbol{U}$ and $\boldsymbol{V}$ yields the following equations to estimate $\boldsymbol{\Lambda}_c$, $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$2 \operatorname{diag}\left(\mathbf{U}\mathbf{U}^\top \boldsymbol{\Lambda}_c \mathbf{V}^\top \mathbf{G}_c \mathbf{V} - \mathbf{U}^\top \mathbf{Z}_\mathbf{c} \mathbf{V}\right) = 0 \qquad c = 1, \ldots, C \tag{9.21a}$$

$$2 \sum_c \mathbf{U} \boldsymbol{\Lambda}_c \mathbf{V}^\top \mathbf{G}_c \mathbf{V} \boldsymbol{\Lambda}_c - 2 \sum_c \mathbf{Z}_c \mathbf{V} \boldsymbol{\Lambda}_c = 0 \tag{9.21b}$$

$$2 \sum_c \mathbf{G}_c \mathbf{V} \boldsymbol{\Lambda}_c \mathbf{U}^\top \mathbf{U} \boldsymbol{\Lambda}_c - 2 \sum_c \boldsymbol{Z}_c^\top \mathbf{U} \boldsymbol{\Lambda}_c = 0 \tag{9.21c}$$

The detailed calculation is given in Appendix B. To the knowledge of the author, a closed-form solution of Eqs. (9.21) does not exist for general $\boldsymbol{\Lambda}_c$, $\boldsymbol{U}$ and $\boldsymbol{V}$. Therefore, a conjugate gradient descent has been used to solve Eq. (9.18) numerically with use of the gradients given in Eqs. (9.21).

So far, the transformations $\boldsymbol{U}$ and $\boldsymbol{V}$ have been tied over all regression classes to simplify the notation. In practice, regression classes as described in Section 9.2.2 have also been used for $\boldsymbol{U}$ and $\boldsymbol{V}$, but with a larger threshold than that used for $\boldsymbol{\Lambda}$:

$$\boldsymbol{A}_c = \boldsymbol{U}_{c'} \, \boldsymbol{\Lambda}_c \, \boldsymbol{V}_{c'}^\top \tag{9.22}$$

Thus, the classes c are a refinement of the classes c' and the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ in Eq. (9.21) become dependent on c'.

The semi-tied MLLR approach has been evaluated on the Verbmobil II task. As initial values for the conjugate gradient descent the results of a singular value decomposition (SVD) of the usual MLLR matrices were used. In the Verbmobil

II test set, a given speaker occurs in several dialogues. To simulate the effect of little adaptation data, the data from the same speaker but different dialogues were not merged for the following tests, i.e. the adaptation was carried out dialog-wise. Therefore the result for the common MLLR approach in Tables 9.3 and 9.4 differs from the respective number in Table 9.2.

As a first test, the transformations $\boldsymbol{U}$ and $\boldsymbol{V}$ were fixed to their initial values (i.e. the results from the SVD of the initial MLLR matrix) and only the values of $\boldsymbol{\Lambda}_c$ were estimated. For that special case, only Eq. (9.21a) has to be solved, which can be solved in closed form for fixed $\boldsymbol{U}$ and $\boldsymbol{V}$ since it is linear in $\boldsymbol{\Lambda}_{c'}$. Thus, the algorithm is as follows:

1. calculate $\boldsymbol{G}_c$ and $\boldsymbol{Z}_c$ according to Eqs. (9.19) and (9.20) from the adaptation data
2. calculate the usual MLLR matrix $\boldsymbol{A}^0_{c'} = \boldsymbol{Z}_{c'}\boldsymbol{G}^{-1}_{c'}$ with $\boldsymbol{Z}_{c'} = \sum_{c \in c'} \boldsymbol{Z}_c$, $\boldsymbol{G}_{c'} = \sum_{c \in c'} \boldsymbol{G}_c$
3. apply SVD to $\boldsymbol{A}^0_{c'}$: $\boldsymbol{A}^0_{c'} = \boldsymbol{U}^0_{c'}\,\boldsymbol{\Lambda}^0_{c'}\,{\boldsymbol{V}^0_{c'}}^\top$
4. calculate $\boldsymbol{\Lambda}_c$ according to Eq. (9.21a)
5. adapt references with $\boldsymbol{A}_c = \boldsymbol{U}^0_{c'}\,\boldsymbol{\Lambda}_c\,{\boldsymbol{V}^0_{c'}}^\top$ and the usual bias term $b_{c'}$: $\hat{\mu}_s = \boldsymbol{A}_c\mu_s + b_{c'}$ for all HMM-states s which belong to regression class c.

The experimental results are given in Table 9.3. The threshold for the transformations (i.e. regression class c') was kept fixed at 6000 observations. This large threshold was chosen to ensure that the resulting matrices can be estimated reliably rather than using the optimal threshold (cf. Table 9.2). It can be seen that there is a clear optimum at 1000 observations or at about 18 regression classes for the diagonal matrix $\boldsymbol{\Lambda}_c$, but the overall improvement over the usual MLLR is very small.

In order to investigate whether the improvement being so small results from fixing the transformations $\boldsymbol{U}$ and $\boldsymbol{V}$, these transformations were reestimated too. As said before, the equation (9.18) has been solved numerically using a conjugate gradient method [Press & Teukolsky$^+$ 02, p. 424]. The algorithm is as follows:

1. calculate $\boldsymbol{G}_c$ and $\boldsymbol{Z}_c$ according to Eqs. (9.19) and (9.20) from the adaptation data
2. calculate the usual MLLR matrix $\boldsymbol{A}^0_{c'} = \boldsymbol{Z}_{c'}\boldsymbol{G}^{-1}_{c'}$ with $\boldsymbol{Z}_{c'} = \sum_{c \in c'} \boldsymbol{Z}_c$, $\boldsymbol{G}_{c'} = \sum_{c \in c'} \boldsymbol{G}_c$
3. apply SVD to $\boldsymbol{A}^0_{c'}$: $\boldsymbol{A}^0_{c'} = \boldsymbol{U}^0_{c'}\,\boldsymbol{\Lambda}^0_{c'}\,{\boldsymbol{V}^0_{c'}}^\top$

Table 9.3: Recognition results for semi-tied MLLR on Verbmobil II. Only $\boldsymbol{\Lambda}_c$ has been estimated, $\boldsymbol{U}_{c'}$ and $\boldsymbol{V}_{c'}$ were kept fixed to the values obtained by a singular value decomposition of conventional MLLR with an observation threshold of 6000. $|C|$ denotes the average number of regression classes, i.e. the average number of adaptation matrices. The corresponding number for $|C'|$ is 3.9 .

			semi-tied MLLR		conventional MLLR
	min. obs.	$\|C\|$	min. obs. $\boldsymbol{U}_{c'}$, $\boldsymbol{V}_{c'}$	WER [%]	WER [%]
baseline	–	0		22.15	22.15
	6000	3.9		20.88	20.88
	100	91.4		21.14	54.10
	500	28.2		20.90	21.98
	1000	18.7	fixed	**20.73**	21.36
	2000	9.4		20.84	20.90
	4000	5.2		20.88	20.68

4. use the values $\boldsymbol{U}_{c'}^0$, $\boldsymbol{\Lambda}_{c'}^0$ and ${\boldsymbol{V}_{c'}^0}^\top$ as start values for the conjugate gradient method with the gradients from Eq. (9.21)

5. update $\boldsymbol{U}_{c'}^i$, $\boldsymbol{\Lambda}_{c'}^i$ and ${\boldsymbol{V}_{c'}^i}^\top$ until the difference of Eq. (9.18) for successive iterations is below a certain threshold.

6. adapt references with $\boldsymbol{A}_c = \boldsymbol{U}_{c'}^i \boldsymbol{\Lambda}_c^i {\boldsymbol{V}_{c'}^i}^\top$ and the usual bias term $b_{c'}$: $\hat{\mu}_s = \boldsymbol{A}_c \mu_s + b_{c'}$ for all HMM-states s which belong to regression class c.

The experimental results are given in Table 9.4. For these experiments, the threshold as described in item 5. of the above algorithm was set to 10^{-5}. For the best result (4000 observations), the experiment was carried out again with a threshold of 10^{-8}. The resulting adaptation matrices differed slightly, but without having an effect on the recognition performance. As can be seen from Table 9.4 the reestimation of the transformation matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ did not improve the recognition performance compared to the results given in Table 9.3.

9.2.4 Discussion

Unfortunately, the approach of semi-tied MLLR matrices has not resulted in an improved speech recognition accuracy yet. The semi-tied approach aims at a refined use of the regression classes. The recognition results shown in Table 9.2 reveals that

Table 9.4: Recognition results for semi-tied MLLR on Verbmobil II. $\mathbf{\Lambda}_c$, $\boldsymbol{U}_{c'}$ and $\boldsymbol{V}_{c'}$ have been estimated. The observation threshold for $\boldsymbol{U}_{c'}$ and $\boldsymbol{V}_{c'}$ was set to 6000. $|C|$ denotes the average number of regression classes, i.e. the average number of adaptation matrices. The corresponding number for $|C'|$ is 3.9 .

			semi-tied MLLR		conventional MLLR
	min. obs.	$\|C\|$	min. obs. $\boldsymbol{U}_{c'}$, $\boldsymbol{V}_{c'}$	WER [%]	WER [%]
baseline	–	0		22.15	22.15
	6000	3.9		20.88	20.88
	100	91.4			54.10
	500	28.2		21.40	21.98
	1000	18.7	6000	20.92	21.36
	2000	9.4		20.88	20.90
	4000	5.2		**20.71**	20.68

the benefit of regression classes in general is rather small compared to the improvement when using only two adaptation matrices, one for the silence models and one matrix for the speech models. The word error rates are 20.74% when using only two matrices and 20.00% when using about 16 matrices on average, which corresponds to a relative reduction in word error rate of only 3.5%. This is rather small compared to the improvements reported in [Leggetter & Woodland 95a] and [Gales 96], though these tests were performed on a small vocabulary task only and used a supervised adaptation scheme. A general problem of the regression class modeling in the RWTH system seems unlikely, as results using a similar regression class modeling are superior than those presented by other groups([Afify & Siohan 00], c.f. Table 9.5) on the same task. Thus further research is needed to get an insight why the improvements obtained by the use of regression classes are only moderate. A first hint can be obtained from [Leggetter & Woodland 95a]: In Table I of that work it is shown that if pooled covariances are assumed (which is called least squares adaptation in [Leggetter & Woodland 95a]), the use of regression classes also gives only a small additional improvement of about 6% rel. reduction in word error rate in comparison to a reduction of about 18% for model specific covariances. On the other hand, the baseline in [Leggetter & Woodland 95a] uses specific covariance matrices, unfortunately it is not stated in that work if mixture-specific or density-specific. Thus assuming a pooled covariance matrix just for the estimation of the MLLR matrix while the acoustic model does use specific covariance matrices may be an inappropriate approximation.

Currently the acoustic model of the RWTH system only uses a globally pooled covariance matrix. Experiments at the RWTH have shown no significant differences for the baseline system using a globally pooled and mixture-specific covariance matrices [Schlüter 02]. However, this result may change if MLLR adaptation is added to the system. The influence of a pooled covariance matrix on the adaptation performance as well as the quality of the approximation of assuming a pooled covariance for estimating the MLLR matrices, while using specific covariance matrices in the acoustic model, has, to the knowledge of the author, not been studied yet and will be a topic for further research.

An additional improvement of about 6% relative reduction in word error rate is reported in [Leggetter & Woodland 95a] if the forward-backward algorithm is applied for the estimation of the MLLR matrix instead of using the Viterbi approximation, which is applied in the RWTH system.

Another interesting outcome of the experiments is that the semi-tied MLLR matrix is in the vast majority of cases superior to the conventional MLLR with the same threshold for the minimum number of observations per regression class. Certainly, the conventional MLLR has much more parameters to be estimated given the same threshold and the conventional MLLR outperforms the semi-tied approach at the respective optimal threshold. But this opens an interesting direction for further research: The semi-tied MLLR may be helpful if only a small amount of adaptation data is available: Assume, enough adaptation data is available to reliably estimate the transformations $\boldsymbol{U}$ and $\boldsymbol{V}$ before the actual recognition test. Than, during the real test, only the diagonal part $\boldsymbol{\Lambda}$ has to be estimated on-line. The combination of the pre-calculated transformations $\boldsymbol{U}$ and $\boldsymbol{V}$ and the on-line estimated diagonal matrix $\boldsymbol{\Lambda}$ may outperform a full transformation matrix calculated previously on the adaptation data because this matrix will not be able to adapt to changes in the transmission channel for instance. On the other hand, the semi-tied approach should outperform diagonal matrices estimated on-line because it offers a more flexible adaptation and additionally may benefit from previously collected adaptation data. The situation just outlined may for instance arise in the following scenario: The speech recognition is used as user access to a voice mail box. When the voice mail box is set up, the user provides the system with a certain amount of speech data for adaptation purposes. The transformations $\boldsymbol{U}$ and $\boldsymbol{V}$ are calculated on this data and the matrix $\boldsymbol{\Lambda}$ on-line every time the user accesses his mail box. This should be superior to estimating a static transformation matrix on the adaptation data because this matrix may not be able to adapt to environmental conditions different from those represented in the adaptation data, for instance if the adaptation data is collected using a conventional phone and the user accesses the mail box via a mobile phone.

9.3 Band-structured MLLR

In Sections 7.4 and 7.5 it has been shown that the VTN warping matrix may be approximated by a band-diagonal structured matrix. Similar approximations have been proposed based on empirical findings in [Uebel & Woodland 99, Afify & Siohan 00]. Although the band-structure has been shown to be valid for the VTN transformation matrix $\boldsymbol{A}_{\text{VTN}}$ in Section 7.4, a similar structure can also been anticipated for the corresponding MLLR matrix. In Section 7.7 VTN has been shown to be equivalent to a C-MLLR with transformation matrix $\boldsymbol{A}_{\text{C-MLLR}} = \boldsymbol{A}_{\text{VTN}}^{-1}$. In [Demko & Moss$^+$ 84] it has been shown that the entries of inverses of band matrices decay fast enough to justify a similar approximation for the inverse matrix itself, though the number of bands may be higher for the inverse matrix. This motivates to restrict the usual MLLR matrix to a band structure.

The experimental setup is as follows: the common MLLR matrix $\boldsymbol{A}$ has been restricted to a band structure: $A_{ij} = 0$ for $j < i - \delta$ and $j > i + \delta$, i.e. only δ bands have been used. Similar results have been reported in [Afify & Siohan 00], where the authors used an iterative approach to solve the occurring equations. Unfortunately, the convergence of the iteration is not guaranteed and in fact convergence problems have been observed by the author using this iterative approach. On the other hand, the estimation of the MLLR matrix is a quadratic optimization problem and thus can be solved in closed form. As the solution for the MLLR matrix can be obtained row-wise, it is easy to restrict the resulting matrix to a band structure. For diagonal covariance matrices

$$\boldsymbol{\Sigma}_s = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2, \ldots, \sigma_D^2) \ ,$$

the solution of Eq. (9.4) for the i-th row $w^{(i)}$ of the band-restricted transformation matrix $\boldsymbol{W}_\delta$ as defined in Eq. (9.2) is given by the following equation:

$$\sum_{k=0,i-\delta}^{i+\delta} G_{jk}^{(i)} w_k^{(i)} = Z_j^{(i)} \quad j = 0, i-\delta, \ldots, i+\delta \tag{9.23}$$

with $\boldsymbol{G}^{(i)}$ and $\boldsymbol{Z}^{(i)}$ from Eqs. (9.7) and (9.8) The $k, j = 0$-term in Eq. (9.23) results from the bias term $b_{s,r}$. Thus, for each row only a part of $\boldsymbol{G}^{(i)}$ of size $(2\delta + 1 + 1^{1}) \times (2\delta + 1 + 1)$ has to be inverted.

The Spoke3 test set of the Wall Street Journal task has been used for the experimental evaluation. This consists of 40 utterances for test and adaptation each, from 10 non-native speakers. Further details of this corpus can be found in Appendix C. The experimental setup [2] is as follows:

- 16 Mel-frequency Cepstral coefficients

[1] for the bias term

[2] The experimental setup has been chosen to obtain comparable results as those presented in [Afify & Siohan 00].

- short-term Cepstral mean normalization
- LDA of seven adjacent augmented MFCC vectors with reduction to 32 dimensions
- 4000 decision-tree clustered, across-word HMM states, one globally pooled diagonal covariance
- 185k Gaussian mixture densities
- 6 state HMM topology
- trigram language model trained on the training data provided by NIST
- 5000 word lexicon
- supervised adaptation estimated on the adaptation utterances, CART tree used for definition of regression classes, observation threshold 100 observations (see Section 9.2.2 for details of the regression class modeling)

The acoustic models were trained on the SI-84 training set. Details of the RWTH speech recognition system can be found in [Sixtus & Ney 02]. The baseline speaker independent (SI) word error rate (WER) is 34.7%, the SI WER on the WSJ0-5k evaluation corpus is 3.75%.

Table 9.5: Recognition results (WER in %) on the WSJ Spoke3 corpus for different amounts of adaptation data and varying number of bands δ for the MLLR matrix.

		Number of adaptation utterances				
		1	5	10	20	40
baseline		34.7				
Number of bands	0	27.7	23.1	20.2	17.3	17.4
	1	27.0	22.2	18.6	16.3	15.5
	2	**27.0**	22.0	17.9	15.6	14.5
	3	27.3	21.3	17.7	14.9	14.0
	4	27.5	**21.2**	16.7	14.1	13.3
	5	28.2	21.9	**16.7**	14.2	13.2
	7	28.8	22.4	16.8	**14.0**	12.9
	full	43.8	33.2	23.2	15.6	**12.5**

As can be seen from Table 9.5 and Fig. 9.2, the largest improvement is obtained from the first few bands of the MLLR matrix. This supports the theoretical findings from Sections 7.4 and 7.5 that the transformation matrix can be approximated by a tri- or quindiagonal matrix. Additionally, the full matrix is superior only for the complete adaptation set of 40 utterances. For all other adaptation sets a band-restricted matrix is superior to the full matrix. In general, except for very limited

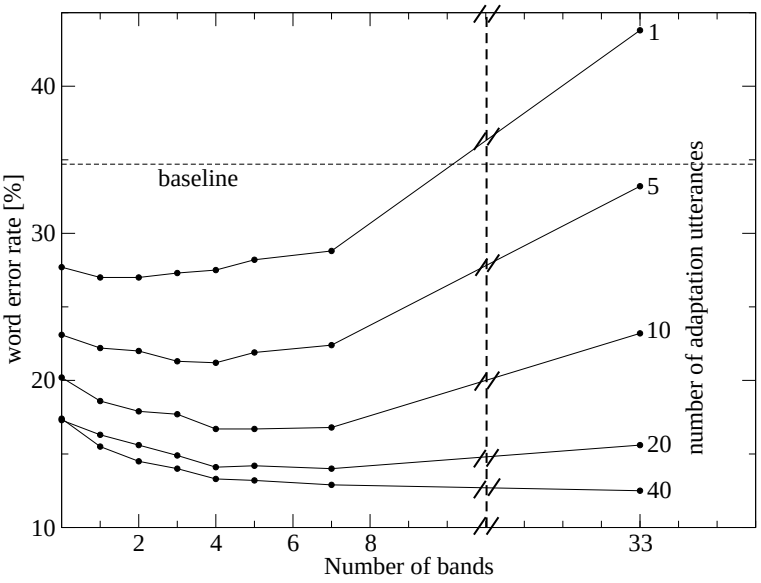

Figure 9.2: Plot of recognition results from Table 9.5.

adaptation data (1 utterance only), the best choice for δ is 4 which is very close to the respective optimal value.

9.4 Confidence Measures

Automatic recognition of conversational speech tends to have significantly higher word error rates (WER) than read speech. Improvements gained from unsupervised speaker adaptation methods like MLLR are reduced because of their sensitivity to recognition errors in the first pass. In this Section it is shown that the use of confidence measures can improve the adaptation performance for conversational speech. Usually, two main adaptation schemes are distinguished: supervised adaptation, where the correct word transcription of the adaptation data is known, and unsupervised adaptation, where it is not known. In unsupervised adaptation a preliminary transcription is generated in a first recognition pass, which usually contains recognition errors. Adaptation with this erroneous transcription degrades the performance of the adaptation step compared to supervised adaptation. Confidence measures can be used to automatically label individual words in the preliminary transcription with a continuous confidence score between 0 and 1. Using a threshold on that confidence score, each word can be labeled either *correct* or *false*, which enables the adaptation step to use only those words which are most probably correct. This is especially significant in the context of conversational speech recognition, as the first pass tends to have higher WER than read speech.

In the following, a short introduction into the specific confidence measure used

in this Section is given. For a more detailed description, the reader is referred to [Wessel 02]. The fundamental rule in all statistical speech recognition systems is Bayes' decision rule (cf. Section 1.1) which is based on the posterior probability $p(w_1^M|x_1^T)$ of a word sequence $w_1^M = w_1, \ldots, w_M$, given a sequence of acoustic observations $x_1^T = x_1, \ldots, x_T$. That word sequence $\left\{w_1^M\right\}_{opt}$ which maximizes this posterior probability also minimizes the probability of an error in the recognized sentence:

$$\begin{aligned}\left\{w_1^M\right\}_{opt} &= \underset{w_1^M}{\text{argmax}} \quad p(w_1^M|x_1^T) \\ &= \underset{w_1^M}{\text{argmax}} \quad \left\{\frac{p(x_1^T|w_1^M)\cdot p(w_1^M)}{p(x_1^T)}\right\} \\ &= \underset{w_1^M}{\text{argmax}} \quad \left\{p(x_1^T|w_1^M)\cdot p(w_1^M)\right\},\end{aligned} \tag{9.24}$$

where $p(w_1^M)$ denotes the language model probability, $p(x_1^T|w_1^M)$ the acoustic model probability and $p(x_1^T)$ the probability of the acoustic observations. Strictly speaking, the maximization is also performed over all sentence lengths M.

If these posterior probabilities were known, the posterior probability $p(w_m|x_1^T)$ for a specific word w_m could easily be estimated by summing up the posterior probabilities of all sentences w_1^M containing this word at position m. This posterior word probability can directly be used as a measure of confidence [Wessel & Macherey$^+$ 98, Wessel & Schlüter$^+$ 00, Wessel & Schlüter$^+$ 01].

Unfortunately, the probability of the sequence of acoustic observations $p(x_1^T)$ is normally omitted since it is invariant to the choice of a particular sequence of words. The decisions during the decoding phase are thus based on unnormalized scores. These scores can be used for a comparison of competing sequences of words, but not for an assessment of the probability that a recognized word is correct. This fact, and in other words the estimation of the probability of the acoustic observations, is the main problem for the computation of confidence measures.

For the following considerations it is very useful to introduce explicit boundaries between the words in a word sequence w_1^M. Let τ denote the starting time and t the ending time of word w. With these definitions, $[w;\tau,t]$ is a specific hypothesis for this word. A sequence of M word hypotheses can thus be formulated as $[w;\tau,t]_1^M = [w_1;\tau_1,t_1],\ldots,[w_M;\tau_M,t_M]$, where $\tau_1 = 1$, $t_M = T$ and $t_{n-1} = \tau_n - 1$ for all $n = 2,\ldots,M$. In order to determine these word boundaries, the following modified Bayes' decision rule is considered. $p([w;\tau,t]_1^M|x_1^T)$ denotes the posterior probability for a sequence of word hypotheses, given the acoustic observations and

$p(x_1^T|[w;\tau,t]_1^M)$ the acoustic model probability:

$$\begin{aligned}\left\{[w;\tau,t]_1^M\right\}_{opt} &= \underset{[w;\tau,t]_1^M}{\operatorname{argmax}} \; p([w;\tau,t]_1^M|x_1^T) \\ &= \underset{[w;\tau,t]_1^M}{\operatorname{argmax}} \; \frac{p(x_1^T|[w;\tau,t]_1^M)\cdot p(w_1^M)}{p(x_1^T)} \\ &= \underset{[w;\tau,t]_1^M}{\operatorname{argmax}} \; \frac{\prod_{m=1}^{M}\left[p(x_{\tau_m}^{t_m}|w_m)\cdot p(w_m|w_1^{m-1})\right]}{p(x_1^T)} \; .\end{aligned} \tag{9.25}$$

It is assumed that the generation of the acoustic observations $x_{\tau_m}^{t_m} = x_{\tau_m},\ldots,x_{t_m}$ depends on word w_m only. With these word boundaries, the posterior probability $p([w;\tau,t]|x_1^T)$ for a specific word hypothesis $[w;\tau,t]$ can be computed by summing up the posterior probabilities of all sentences which contain the hypothesis $[w;\tau,t]$:

$$\begin{aligned}p([w;\tau,t]|x_1^T) &= \\ &= \sum_{\substack{[w;\tau,t]_1^M:\\ \exists\, n\in\{1,\ldots,M\}:\\ [w_n;\tau_n,t_n]=[w;\tau,t]}} \frac{\prod_{m=1}^{M}\left[p(x_{\tau_m}^{t_m}|w_m)\cdot p(w_m|w_1^{m-1})\right]}{p(x_1^T)} \; .\end{aligned} \tag{9.26}$$

The posterior probability for a word hypothesis and $p(x_1^T)$ can be computed on the basis of word graphs. In the style of the forward-backward algorithm the forward probability and the backward probability for a word hypothesis are computed and both probabilities are combined into the posterior probability of this hypothesis. In contrast to the forward-backward algorithm on a Hidden-Markov-Model state level, the forward-backward algorithm is now based on a word hypothesis level.

These posterior hypothesis probabilities turned out to perform poorly as a confidence measure. In fact, this observation is not surprising since the fixed starting and ending time of a word hypothesis determine which paths in the word graph are considered during the computation of the forward-backward probabilities. Usually, several hypotheses with slightly different starting and ending times represent the same word and the probability mass of the word is split among them. In order to solve this problem, the posterior probabilities of all those hypotheses which represent the same word have to be summed up. More details are given in [Wessel & Schlüter$^+$ 01, Wessel 02].

Using Confidence Measures for MLLR Adaptation

For an unsupervised two-pass adaptation strategy the accuracy of the first pass is crucial for the adaptation performance. Recognition errors and out-of-vocabulary words degrade the adaptation to the new speaker or environment. The use of confidence measures for unsupervised speaker adaptation has been investigated before in [Anastasakos & Balakrishnan 98, Nguyen & Gelin$^+$ 99]. The improvements reported by these authors range from 1.8% to 4.1% relative reduction in word error rate when comparing MLLR adaptation with and without confidence measures.

To study the effect of confidence measures on the MLLR adaptation in the RWTH system, the word posterior based confidence measures described above were applied as follows: each word in the recognized sentence was annotated with a confidence score between 0 and 1. When collecting the statistics according to Eqs. (9.12) and (9.13), those feature and mean vectors belonging to a word with a confidence score below a given threshold were omitted. In other words, all words in the preliminary transcription with a bad confidence score were disregarded for the estimation of the MLLR adaptation matrix. The experimental results are given in Table 9.6. The experimental setup is as follows:

- 16 Mel-frequency Cepstral coefficients
- short-term Cepstral mean and variance normalization on a sliding window of 2 seconds length
- first order derivatives (using linear regression on five frames), second order derivative of energy
- LDA of three adjacent augmented MFCC vectors with reduction to 33 dimensions
- 2500 decision-tree clustered, within-word HMM states
- 243k Gaussian mixture densities
- 3 state HMM topology
- 10000 word lexicon
- two-pass unsupervised adaptation using 2 adaptation matrices and 50 bias vectors as described in Section 9.2.1

For this hard decision scheme a confidence threshold of 0.6 was applied and an improvement in word error rate of 4.9% rel. could be achieved. Varying the threshold in the range from 0.4 to 0.7 yielded no significant difference. Similar relative improvements were reported in [Wallhoff & Willet$^+$ 00], however only on clean, read speech (WSJ0 task) and with a relatively high baseline word error rate for this task.

A drawback of completely disregarding the time frames with bad confidence scores is the reduction in the amount of adaptation data. Thus, in another experiment, the

Table 9.6: Recognition test results on Verbmobil II using confidence measures for MLLR adaptation.

adaptation method	WER[%]
no adaptation (baseline)	24.6
without confidence measures	22.6
with confidence measures, hard decision	**21.5**
with confidence measures, weighted	22.0
only correctly recognized words	21.0
supervised	18.3

feature and mean vectors were weighted with the confidence score annotated to the corresponding word rather than disregarding bad time frames completely. Although the amount of adaptation data is about 20% larger, the recognition performance is inferior to the hard decision scheme. This suggests that the amount of adaptation data is already sufficient when using only those time frames with a good confidence score. On the other hand, the contribution of correctly recognized words is scaled down when weighted with the confidence score.

In order to get an insight in the best performance possible with confidence measures, the statistics in Eqs. (9.12) and (9.13) were collected using only the correctly recognized words. This is equal to an "ideal" confidence measure with 0% false acceptance and 0% false rejection rate. As can be seen from Table 9.6, the result obtained with word posterior based confidence measures is already quite close to the ideal confidence measure. Another interesting contrast experiment is to use the correct transcription of the test data in the first pass for a supervised adaptation. Evidently, this is not a feasible experiment for real applications but gives more insight in the potential performance of MLLR. This experiment gives an upper bound for possible improvements to be gained by MLLR adaptation for a given set of adaptation classes. Comparing the results obtained with the correctly recognized words only and with the complete correct transcription reveals still a major difference in recognition performance. This shows the importance of adapting especially to those words that were incorrectly recognized in the first pass and thus reduce the mismatch between the acoustic models and the acoustic vectors of those words which are responsible for recognition errors.

9.5 MLLR in Combination With Other Adaptation Approaches

This Section deals with the combination of the MLLR adaptation with other adaptation approaches, namely Vocal Tract Normalization (VTN) and Quantile Equalization. The goal is to evaluate to which extend the gains obtained by each method individually are additive when the approaches are combined.

9.5.1 MLLR and VTN

In [Pye & Woodland 97] and [Uebel & Woodland 99] it has been shown that improvements obtained by VTN and MLLR are largely additive, whereas in [Uebel & Woodland 99] it has been reported that there is only a very small benefit from combining VTN and C-MLLR (constrained MLLR, i.e. mean and variances are adapted using the same transformation matrix). An explanation of that empirical finding has been given in Chapter 7, where it is shown that VTN can be expressed as a special case of C-MLLR. However, there are benefits of combining VTN and (standard) MLLR, experimental results are given in Table 9.7.

Table 9.7: Recognition test results on Verbmobil II using VTN and MLLR adaptation. The results indicated by $(*)$ and $(**)$ are preliminary, see text for experimental details and further explanation.

	MLLR	VTN	WER[%]
Baseline	no	no	23.7
	yes	no	22.8 $(*)$
	no	yes	21.3 $(**)$
	yes	yes	19.8 $(**)$

The experimental setup use in the experiments above is as follows:

- 16 Mel-frequency Cepstral coefficients
- short-term Cepstral mean and variance normalization on a sliding window of 2 seconds length
- LDA of eleven adjacent MFCC vectors (no derivatives) with reduction to 33 dimensions
- 2500 decision-tree clustered, within-word HMM states
- 286k Gaussian mixture densities
- 3 state HMM topology

- 10000 word lexicon
- two-pass unsupervised adaptation in the case of MLLR
- supervised estimation of the VTN warping factors

These numbers indicated by (∗) and (∗∗) are preliminary results. At the time these experiments were carried out the RWTH system was redesigned and newly implemented. The MLLR had been implemented already in the new system, whereas the estimation of the warping factors for VTN had not yet been implemented. Therefore the VTN warping factors used in the experiment indicated by (∗∗) in Table 9.7 were estimated in a supervised scheme using the previous system. This results in the very good performance of VTN compared to the MLLR adaptation. Additionally, the performance of the MLLR itself in the within-word system is less compared to the performance in the across-word model system (WER 22.2% → 20.0%, cf. Table 9.2). Presumably, there has been an error in the time alignment part of the new RWTH software for within-word modeling, which is supported the the observation that other approaches which rely on that module suffer from reduced performance, too [Zolnay 04]. The problem could not be solved completely during the scope of this work.

Apart from the degraded performance of MLLR for within-word modeling, the combination of VTN and MLLR is mostly additive, as has also been shown in [Uebel & Woodland 99].

9.5.2 MLLR and Quantile Equalization

Quantile equalization (QE) [Hilger & Ney 01] aims at increasing the noise robustness of the automatic speech recognition system. The basic idea is to transform the feature vectors during the signal analysis based on the cumulative distribution function. The cumulative distribution function for the training and testing data set will differ if a mismatch of the training and testing conditions exists. The mismatch can be reduced by a non-parametric approach (*histogram normalization*, [Molau & Keysers$^+$ 02]). This approach requires the estimation of the cumulative distribution of the test data and thus a large number of test data for a reliable estimation. In order to provide an online capable approach, the cumulative distribution is approximated by a small number of quantiles. The cumulative distributions of the training and test data are now matched based on the quantiles using a non-linear parametric transformation function rather than the full histograms. The transformation function T_k is applied to the kth Mel-scaled filter bank coefficient Y_k before taking the logarithm (i.e. each filter bank coefficient is transformed individually), yielding the transformed coefficient $\tilde{Y_k}$:

$$\tilde{Y_k} = T_k(Y_k, \theta_k) = Q_{kN_Q}\left(\alpha_k\left(\frac{Y_k}{Q_{kN_Q}}\right)^{\gamma_k} + (1-\alpha_k)\frac{Y_k}{Q_{kN_Q}}\right) \tag{9.27}$$

The parameters $\theta_k = \{\alpha_k, \gamma_k\}$ of the transformation are estimated by optimizing the squared distance between the transformed recognition quantiles and the training quantiles:

$$\theta_k = \underset{\theta'_k}{\operatorname{argmin}} \left(\sum_{i=1}^{N_Q-1} \left(T_k(Q_{ki}, \theta'_k) - Q_i^{train} \right)^2 \right) \tag{9.28}$$

The QE approach can be further enhanced by combining neighboring filter outputs [Hilger & Ney$^+$ 03]: A second linear transformation is applied to the quantiles $\tilde{Q}_k$ which are obtained from the transformed values $\tilde{Y}_k$ after the transformation of Eq. (9.27):

$$\tilde{T}_k(\tilde{Q}_k) = (1 - \lambda_k - \rho_k)\tilde{Q}_k + \lambda_k \tilde{Q}_{k-1} + \rho_k \tilde{Q}_{k+1} \quad . \tag{9.29}$$

Again, the parameters $\{\lambda_k, \rho_k\}$ are obtained by minimizing the squared distance

$$\{\lambda_k, \rho_k\} = \underset{\{\lambda'_k, \rho'_k\}}{\operatorname{argmin}} \left(\sum_{i=1}^{N_Q-1} \left(\tilde{T}_k(\tilde{Q}_{k,i}) - Q_i^{train} \right)^2 + \beta \left(\lambda'^2_k + \rho'^2_k \right) \right) \quad , \tag{9.30}$$

where β is a scaling factor to restrict the possible values of λ_k and ρ_k.

In addition, using the 10th root instead of the logarithm to compress the spectral dynamics has emerged to reduce the word error rate significantly for noisy data. For further details on QE, filter combination and using the 10th root, the reader is referred to [Hilger 04].

The following experiments shall explore to which amount the improvements gained by MLLR and QE individually add up if both approaches are combined. The experimental setup is as follows:

- 16 Mel-frequency Cepstral coefficients with filter mean normalization and 10th root instead of logarithm
- LDA of seven adjacent MFCC vectors (no derivatives) with reduction to 33 dimensions
- 4001 decision-tree clustered, across-word HMM states
- 220k Gaussian mixture densities
- 6 state HMM topology
- 5000 word lexicon
- two-pass unsupervised adaptation in the case of MLLR, CART tree used for definition of regression classes, observation threshold 2000 observations
- utterance wise QE

Recognition test results on the Aurora 4 noisy WSJ 16kHz test set [Hirsch 02] are given in Table 9.8. The Aurora 4 test set consists of 14 different test sets, the given results are averaged over all test sets.

The MLLR approach alone (22.2% WER) outperforms the best QE result (25.5% WER). However, the QE approach can be applied in an online system whereas the MLLR is applied in a two-pass way. The improvements obtained from MLLR adaptation are consistent over all test sets (see [Hilger 04] for details). Test set 8 (clean data, no additional noise, only microphone mismatch) is outstanding: The WER using QE with filter combination is 21.2%, a combination with MLLR yield a WER of 10.3%. Evidently, MLLR is very well capable of compensating for mismatch conditions that are constant over the speaker sessions. On the other hand, QE is very well suited for rapidly changing noise conditions. Thus a combination of MLLR and QE yields significant improvements over either method alone.

Table 9.8: Recognition test results on the Aurora 4 noisy WSJ 16kHz test set using QE and MLLR. QE: standard quantile equalization, QEF: quantile equalization with filter combination.

	QE	MLLR	WER[%]
Baseline	no	no	29.7
	QE	no	25.9
	QEF	no	25.5
	no	yes	22.2
	QE	yes	20.4
	QEF	yes	20.1

9.6 Summary

In this Chapter several enhancements of maximum likelihood linear regression (MLLR) adaptation have been presented. By the use of a refined regression class modeling for the bias vectors of the affine transformation the exploitation of the adaptation data could be improved, which resulted in an improved recognition accuracy. The adaptation data could be even better exploited by the use of a dynamic selection of regression classes. A new approach called semi-tied MLLR has been presented, which aims at an even more flexible definition of regression classes. Although this approach has not resulted in an improved recognition accuracy yet, several promising ways of further research have opened up and should result in an improved recognition accuracy especially for the case of limited available adaptation data.

The restriction of the MLLR to a band structured matrix motivated by the structure of the matrix for vocal tract normalization derived in Chapter 7 resulted in a

significant improvement in recognition accuracy when only limited adaptation data are available.

For conversational speech the adaptation performance usually suffers from higher error rates of the first pass in an unsupervised two-pass adaptation approach. Confidence measures based on word posterior probabilities have shown to improve the adaptation performance for conversational speech. The improvements obtained by computed confidence measures have been shown to be very close to those obtained when using an "ideal" confidence measure where only the correctly recognized words have been used for adaptation.

Finally it has been shown that MLLR is very well suited for a combination with VTN and QE, as for both cases the improvements obtained by the individual methods have shown to be largely additive to the improvements obtained by the MLLR approach.

Chapter 10

Scientific Contributions

The aim of this thesis was to study the interrelationship between several linear transformations which had been commonly used in state-of-the-art speech recognition systems. Due to this popularity, a confusing abundance of linear transformations at various stages of the speech recognition process has been proposed in literature. Accordingly, there exist strong relationships or even equivalences between those transformations. Although some indications of these relationships have been reported, there has been no detailed analysis and classification of the approaches.

In this work a formal description of speaker adaptation techniques, usually referred to as transformation of the model parameters, and speaker normalization, usually referred to as transformation of the acoustic feature vectors, was given. Both speaker adaptation and normalization were described in the same mathematical framework and were deduced to be equivalent in terms of the Bayes' decision rule.

Motivated by this equivalence, a review of the most commonly used linear transformation was given in a unified notation. Many close relationships between the transformations were revealed, for instance between the various speaker adaptation techniques based on linear transformations and vocal tract normalization (VTN). Further close relationships were proven to exist between linear discriminant analysis (LDA), semi-tied covariances (STC), (extended) maximum likelihood linear transformations ((E)MLLT) and heteroscedastic discriminant analysis (HDA). It was derived that all these transformations can be obtained by maximizing a common optimization function for the case of Gaussian emission probabilities.

An improved approach was presented to calculate Mel frequency Cepstral coefficients (MFCC) from the acoustic data. This new approach avoids the twofold smoothing inherent in the common MFCC computation scheme by means of the filter bank and the subsequent reduction of Cepstral coefficients. The Cepstral coefficients are directly calculated from the Fourier spectrum, thus omitting the filter bank completely. This allows for a very compact implementation of the signal analysis front end. Additionally, instead of optimizing a set of independent parameters (number and shape of the filter banks and the number of Cepstral coefficients), only one single parameter has to be optimized for a new task. Additionally, all types of frequency warping can be directly integrated into the Cepstrum transformation

without quantization and interpolation errors.

Based on this modified signal analysis, it was possible to proof analytically that vocal tract normalization is a linear transformation in the Cepstral space for arbitrary invertible warping functions. For three widespread warping functions the transformation matrix was calculated explicitly. Since typical warping functions deviate only slightly from unity, the structure of the matrix describing the linear transformation is determined by general characteristics of the warping functions rather than the actual functional form. It was proven that the transformation matrix is dominated by the main diagonal and a few off-diagonal bands, independent of the actual functional form of the warping function.

Using this general structure to constrain the common maximum likelihood linear regression (MLLR) matrix to a band structure significantly improved the MLLR performance for small amounts of adaptation data. As predicted from the theoretical analysis, increasing the number of off-diagonals the largest improvements are obtained by the first few bands. When using only five adaptation utterances, the common MLLR gave an relative improvement of 4% while using a diagonal MLLR matrix improved the recognition by 33% relative and a band-diagonal MLLR with 4 bands by 39%.

Since vocal tract normalization was expressed as a linear transformation, the Jacobian determinant of that transformation was calculated for the first time for any warping function. Up to now, the Jacobian determinant had either been been simply omitted or taken into account only for a few special warping functions. An experimental study revealed that the consideration of the Jacobian determinant had a significant influence on the estimation of the warping factors but surprisingly not on the recognition performance. Thus the current practice of omitting the Jacobian determinant was confirmed.

Confidence measures based on word posterior probabilities significantly improved the MLLR adaptation performance. With help of these confidence measures the difficulty of recognition errors in the first pass of an unsupervised adaptation scheme had been largely overcome. The use of confidence measures on an conversational speech task improved the recognition accuracy by 4.9% rel. A control experiment where only the correctly recognized words were used for adaptation (thus providing an ideal confidence measure) revealed that this result is already quite close to the optimal value.

Several new approaches were presented to deal with the problem of using adaptation classes for MLLR adaptation when only small amounts of adaptation data are available. First, the use of more refined regression classes for the offset term of the affine transformation was suggested. The performance of the MLLR adaptation was improved by 8.8% relative compared to the baseline or by 2.6% rel. compared to standard MLLR, respectively. By a dynamic selection and a tree organization of the adaptation classes the MLLR yielded an improvement of 9.7% rel. with respect to the baseline. To exploit the adaptation data even more, a subspace structuring

for MLLR adaptation, called semi-tied MLLR, was suggested. The goal is to define a common subspace in which an approximation of diagonal MLLR matrices for the different adaptation classes is most appropriate. The estimation formula were derived and first experimental results were presented. While semi-tied MLLR proved to be inferior to standard MLLR with the same number of regression classes, the overall performance of MLLR could be improved only slightly. However, for the case of very small amounts of adaptation data semi-tied MLLR may improve the recognition performance whereas the standard MLLR actually decreases the recognition performance.

Chapter 11

Outlook

In this thesis linear transformations, especially for speaker adaptation and normalization have been studied. It has been shown that Vocal Tract Normalization (VTN) can be expressed as linear transformation in the Cepstral space and thus VTN emerges as special case of Maximum Likelihood Linear Transformation (MLLR). A study on commonly used linear transformations has demonstrated further close relationships among those transformations. Finally, several enhancements of the MLLR adaptation technique have been developed.

The following questions remain open and may serve as starting point for further research:

- The influence of the Jacobian determinant on the warping factor estimation for VTN warping has been studied in detail in this work. The correct consideration of the Jacobian determinant has a significant effect on the estimated warping factors, which results in a major difference in the warping factor histograms for estimation schemes that both consider and neglect the Jacobian determinant. However, the recognition performance remains almost unaffected by considering or neglecting the Jacobian determinant. This result is even more surprising since the recognition performance is usually quite sensible to changes in the warping factor histograms. So there remains a need for further research on the effects of the Jacobian determinant on the warping factor estimation. A study on different warping factor estimation schemes would give further interesting insights. The scheme used in the RWTH speech recognition system is quite stable. However, the correct consideration of the Jacobian determinant could for instance improve an iterative warping factor estimation, which usually becomes unstable after several iterations. Here the Jacobian determinant may be helpful for an improved warping factor estimation.

- A general structure of the VTN warping matrix has been derived that is almost independent of the specific functional form of the VTN warping function. This structure has been successfully applied for MLLR adaptation and resulted in significant improvements for small amounts of adaptation data. A very promising extension is to use this general structure for speaker adaptive training with

constrained MLLR. Although mathematically more difficult because of the numerical computation of the adaptation matrix, restricting the matrix to a band structure should result in an improved adaptation performance if only limited data from a specific speaker are available for adaptation. In speaker adaptive training often clusters of training speakers are build to ensure that sufficient training data are available to reliably estimate the adaptation matrices. The restriction of the adaptation matrix to a band structure could result in more refined clusters and thus improve the adaptation performance.

- No systematic study on the influence of variance modeling on MLLR adaptation has been published to this date. In this work the benefit from regression classes was less than expected, which could be associated with the pooled covariance modeling in the RWTH system. For the RWTH baseline system, a study on the covariance modeling and the difference between Baum-Welsh and Viterbi training revealed no significant differences. This may change if MLLR speaker adaptation is applied. So the investigation should be repeated in the light of MLLR speaker adaptation.

- The semi-tied MLLR modeling proposed in this work gave very promising results, although the overall recognition performance could not be improved. An interesting direction of further research is to estimate the transformations $\boldsymbol{U}$ and $\boldsymbol{V}$ of the semi-tied MLLR approach beforehand, either on adaptation data from the actual speaker collected in advance or on a training speaker with similar characteristics as the current test speaker. For example, text-independent Gaussian mixture models may be trained on the training speakers or clusters of training speakers. In recognition, the most probable training speaker according to the Gaussian mixture models could be chosen, for instance based on the acoustic vectors of the first sentence. Then the transformations $\boldsymbol{U}$ and $\boldsymbol{V}$ of that training speaker are used for semi-tied MLLR modeling and only the diagonal matrix $\boldsymbol{\Lambda}$ has to be estimated online for the given test speaker. Thus a full transformation matrix may be used to adapt the acoustic model while only a diagonal transformation matrix has to be estimated on the test data. This approach should result in an improved adaptation performance, for instance for fast adaptation, where the adaptation matrices are estimated iteratively on the test data.

Appendix A

Symbols and Acronyms

In this appendix, all relevant mathematical symbols and acronyms which are used in this thesis are defined for convenience. Detailed explanations are given in the corresponding Chapters.

Mathematical Symbols

In general, mathematical symbols printed in **bold** face denote matrices, $\boldsymbol{A}^\top$ denotes a transposed matrix.

x_1^T	sequence of acoustic vectors $x_1, \ldots, x_T$
T	total number of time frames
w_1^N	word sequence $w_1, \ldots, w^N$
N	total number of spoken words
$p(x_1^T \vert w_1^N)$	acoustic probability distribution of the acoustic vector sequence x_1^T given the word sequence w_1^N (acoustic model)
$p(w_1^N)$	a-priori probability of the word sequence w_1^N (language model)
$s = 1, \ldots, S$	index for hidden Markov model states
$r = 1, \ldots, R$	index for different speakers
s_1^T	sequence of hidden Markov model states $s_1, \ldots, s_T$, usually the Viterbi path
$\mathcal{N}(x \vert \mu, \boldsymbol{\Sigma})$	Gaussian probability distribution
μ	mean vector, usually of an Gaussian probability distribution
$\boldsymbol{\Sigma}$	covariance matrix, usually of an Gaussian probability distribution

μ_s	mean vector of HMM emission probability of state s
$\mathbf{\Sigma}_s$	covariance matrix of HMM emission probability of state s
c_{sl}	mixture weight for density l of state s, usually of an Gaussian mixture distribution
θ	set of model parameters, usually $\{\{\mu\}, \{c\}, \mathbf{\Sigma}\}$
$\mathrm{X}_{\mathrm{Train}}$	set of training data, consisting of a collection of different acoustic conditions
$\mathrm{X}_{\mathrm{Test}}$	test data, usually consisting of only one specific condition
$\tilde{\mathrm{X}}$	normalized acoustic data
θ_{Train}	model parameter set trained on $\mathrm{X}_{\mathrm{Train}}$, covers the conditions represented in $\mathrm{X}_{\mathrm{Train}}$
θ_{Test}	(hypothetical) model parameter set that covers the specific test conditions represented in $\mathrm{X}_{\mathrm{Test}}$
$\tilde{\theta}$	model parameter set trained on the normalized acoustic data $\tilde{\mathrm{X}}$
ω	frequency
$\tilde{\omega}$	warped frequency
ω_0	inflexion point for piece-wise linear VTN warping
$\tilde{\omega}_0$	warped inflexion point for piece-wise linear VTN warping
ω_{mel}	mel-frequency
$\tilde{\omega}_{\mathrm{mel}}$	warped mel-frequency
α	warping factor for vocal tract normalization (VTN)
g_α	warping function with warping factor α
$\boldsymbol{A}_{s,r}$	maximum likelihood linear regression (MLLR) transformation matrix for state s of speaker r
x_{r1}^{T}	sequence of acoustic vectors for speaker r, adaptation data
w_{r1}^{N}	sequence of spoken words from speaker r, transcription of adaptation data
$b_{s,r}$	MLLR bias vector for state s of speaker r

$\boldsymbol{H}$	transformation matrix for the covariance matrix
ξ_s	extended mean vector $\xi_s = [1\ \mu_s^\top]^\top$ for MLLR adaptation
$\boldsymbol{W}_{s,r}$	extended MLLR matrix $[b_{s,r}\ \boldsymbol{A}_{s,r}]$
$\gamma_s(t)$	state occupation probability for state s
σ_{si}	i-th diagonal element of a diagonal covariance matrix
D	dimension of the acoustic feature space
$\boldsymbol{\Theta}$	HDA transformation matrix
$\boldsymbol{\Theta}^{-\top}$	$\boldsymbol{\Theta^{-1}}^\top$
c_k	k-th Cepstral coefficient
A_{nk}	nth row, kth column element of VTN warping matrix
δ_{nk}	Kronecker delta, $\delta_{nk} = 1$ for $n = k$; 0 otherwise
$\circ$	composition of two functions, $(f \circ g)(x) = f(g(x))$
$X(\omega)$	Fourier spectrum
$\tilde{X}(\tilde{\omega})$	warped Fourier spectrum
s_k	symmetry factor, $s_k = \frac{1}{2}$ for $k = 0$; 1 otherwise
$S(X)$	Fresnel sine function
$C(X)$	Fresnel cosine function
$g^{(-1)}$	inverse warping function
$\mathrm{tr}(\boldsymbol{A})$	trace of matrix $\boldsymbol{A}$
$\mathrm{diag}(\boldsymbol{A})$	diagonal elements of $\boldsymbol{A}$
$\det \boldsymbol{A}$	determinant of $\boldsymbol{A}$

Acronyms and Abbreviations

Acronyms

ASR	**a**utomatic **s**peech **r**ecognition
BLT	**bi**llinear **t**ransform
C-MLLR	**c**onstrained **m**aximum **l**ikelihood **l**inear **r**egression
CART	**c**lassification **a**nd **r**egression **t**ree
CC	**C**epstral **c**oefficient(s)
CMN	**C**epstral **m**ean **n**ormalization
(D)ARPA	(**D**efense) **A**dvanced **R**esearch **P**rojects **A**gency
DCT	**d**iscrete **c**osine **t**ransform
DLLR	**d**iscounted **l**ikelihood **l**inear **r**egression
EM	**e**xpectation **m**aximization
EMLLT	**e**xtended **m**aximum **l**ikelihood **l**inear **t**ransformation
F-MLLR	**f**eature-space **m**aximum **l**ikelihood **l**inear **r**egression
FFT	**f**ast **F**ourier **t**ransform
HDA	**h**eteroscedastic **d**iscriminant **a**nalysis
HMM	**h**idden **M**arkov **m**odel
LDA	**l**inear **d**iscriminant **a**nalysis
LM	**l**anguage **m**odel
LPC	**l**inear **p**redictive **c**oding
LVCSR	**l**arge **v**ocabulary **c**ontinuous **s**peech **r**ecognition
MAP	**m**aximum **a**-**p**osteriori
MAPLR	**m**aximum **a**-**p**osteriori **l**inear **r**egression
MFCC	**M**el-**f**requency **C**epstral **c**oefficients
MF-PLP	**M**el **f**requency **p**erceptual **l**inear **p**rediction

ML	**m**aximum **l**ikelihood
MLLR	**m**aximum **l**ikelihood **l**inear **r**egression
MLLT	**m**aximum **l**ikelihood **l**inear **t**ransformation
NAB	**N**orth **A**merican **B**usiness **N**ews - a speech corpus
PLP	**p**erceptual **l**inear **p**rediction
PP	**p**er**p**lexity
RWTH	**R**heinisch-**W**estfälische **T**echnische **H**ochschule
SAT	**s**peaker **a**daptive **t**raining
SI	**s**peaker **i**ndependent
SIMD	**s**ingle **i**nstruction, **m**ultiple **d**ata
STC	**s**emi-**t**ied **c**ovariance modeling
SVD	**s**ingular **v**alue **d**ecomposition
TI-DIGITS	**T**exas **I**nstruments connected **digit** **s**equences - a speech corpus
VTN	**v**ocal **t**ract length **n**ormalization
WER	**w**ord **e**rror **r**ate
WSJ	**W**all **S**treet **J**ournal - a speech corpus

Appendix B

Detailed Calculations

B.1 Semi-tied MLLR modeling

This section gives come detailed calculations for semi-tied MLLR modeling, which has been discussed in Section 9.2.3. For semi-tied MLLR, the optimization problem

$$\underset{\{\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{V}\}}{\operatorname{argmin}} \sum_{t=1}^{T} \|x_t - \boldsymbol{A}_c \mu_{s_t}\|^2 \tag{9.18}$$

with

$$\boldsymbol{A}_c = \boldsymbol{U}\,\boldsymbol{\Lambda}_c\,\boldsymbol{V}^\top \qquad \text{and} \qquad \boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_c \mid c = 1, \ldots, C\}\ .$$

and $\boldsymbol{G}_c$ and $\boldsymbol{Z}_c$ given as

$$\boldsymbol{G}_c = \sum_{\substack{t=1 \\ s_t \in c}}^{T} \mu_{s_t} \mu_{s_t}^\top \tag{9.19}$$

$$\boldsymbol{Z}_c = \sum_{\substack{t=1 \\ s_t \in c}}^{T} x_{rt} \mu_{s_t}^\top \tag{9.20}$$

has to be solved. For convenience, the function to be optimized is defined as

$$F(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{V}) := \sum_{t=1}^{T} \|x_t - \boldsymbol{A}_c \mu_{s_t}\|^2 = \sum_{t=1}^{T} \sum_{d=1}^{D} \left(x_{td} - (\boldsymbol{U}\,\boldsymbol{\Lambda}_c\,\boldsymbol{V}^\top \mu_{s_t})_d\right)^2 \tag{B.1}$$

where x_{td} denotes the dth component of the D-dimensional vector x_t; accordingly $(\bullet)_d$ denotes the d-th component of the term in parentheses, the symbol $\bullet$ denotes the term in parentheses itself. The derivatives will be calculated component-wise.

B.1.1 Derivative w.r.t. $\mathbf{\Lambda}_c$

As $\mathbf{\Lambda}_c$ is modeled to be diagonal, only the diagonal components have to be calculated:

$$\frac{\partial}{\partial \Lambda_{cii}} F(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{V}) =$$

$$= -2 \sum_{t=1}^{T} \sum_{d=1}^{D} \left(x_{td} - (\boldsymbol{U}\, \boldsymbol{\Lambda}_{c'}\, \boldsymbol{V}^\top \mu_{s_t})_d \right) \cdot \frac{\partial}{\partial \boldsymbol{\Lambda}_{\boldsymbol{c'}\, ii}} (\boldsymbol{U}\, \boldsymbol{\Lambda}_{c'}\, \boldsymbol{V}^\top \mu_{s_t})_d \tag{B.2a}$$

$$= -2 \sum_{t=1}^{T} \sum_{d=1}^{D} (\bullet)\, \frac{\partial}{\partial \boldsymbol{\Lambda}_{\boldsymbol{c}ii}} \sum_{m,n=1}^{D} U_{dm}\, \Lambda_{c'\, mm}\, V_{nm}\, \mu_{s_t\, n} \tag{B.2b}$$

$$= -2 \sum_{c'=1}^{C} \sum_{t \in c'} \sum_{d=1}^{D} (\bullet) \sum_{m,n=1}^{D} U_{dm}\, V_{nm}\, \mu_{s_t\, n}\, \delta_{im} \delta_{cc'} \tag{B.2c}$$

$$= -2 \sum_{t \in c} \sum_{d=1}^{D} \left(x_{td} - (\boldsymbol{U}\, \boldsymbol{\Lambda}_{c}\, \boldsymbol{V}^\top \mu_{s_t})_d \right) \sum_{n=1}^{D} U_{di}\, V_{ni}\, \mu_{s_t\, n} \tag{B.2d}$$

$$= 2 \sum_{d,n,k,l=1}^{D} U_{dk}\, \Lambda_{kk}\, V_{lk}\, U_{di}\, V_{ni} \sum_{t \in c} \mu_{s_t\, l}\, \mu_{s_t\, n} - 2 \sum_{d,n=1}^{D} U_{di} V_{ni} \sum_{t \in c} x_{td}\, \mu_{s_t\, n} \tag{B.2e}$$

$$= 2 \sum_{d,n,k,l=1}^{D} U_{id}^\top U_{dk}\, \Lambda_{kk}\, V_{kl}^\top\, G_{cnl} V_{ni} - 2 \sum_{d,n=1}^{D} U_{id}^\top Z_{cdn} V_{ni} \tag{B.2f}$$

$$= 2\, \mathrm{diag}\left(\mathbf{U}\mathbf{U}^\top \mathbf{\Lambda}_c \mathbf{V}^\top \mathbf{G}_c \mathbf{V} - \mathbf{U}^\top \mathbf{Z}_{\mathbf{c}} \mathbf{V} \right) \tag{B.2g}$$

In Eq. (B.2c) the Viterbi approximation is utilized. The meaning of the shortcut $t \in c$ is as follows: The sum is taken over all time frames for that the Viterbi state s_t (and thus the mean μ_{s_t}) belongs to class c.

B.1.2 Derivative w.r.t. U_{ij}

$$\frac{\partial}{\partial \boldsymbol{U}_{ij}} F(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{V}) =$$

$$= -2 \sum_{c=1}^{C} \sum_{t \in c} \sum_{d=1}^{D} \left(x_{td} - (\boldsymbol{U}\, \boldsymbol{\Lambda}_c\, \boldsymbol{V}^\top \mu_{s_t})_d \right) \cdot \sum_{l=1}^{D} \Lambda_{cjj}\, V_{lj}\, \mu_{s_t l}\, \delta_{id} \tag{B.3a}$$

$$= -2 \sum_{c=1}^{C} \sum_{d,l=1}^{D} \Lambda_{cjj}\, V_{lj}\, Z_{cdl}\, \delta_{id} + 2 \sum_{c=1}^{C} \sum_{t \in c} \sum_{d,l=1}^{D} (\boldsymbol{U}\, \boldsymbol{\Lambda}_c\, \boldsymbol{V}^\top \mu_{s_t})_d\, \Lambda_{cjj}\, V_{lj} \mu_{s_t l}\, \delta_{id} \tag{B.3b}$$

$$= -2 \sum_{c=1}^{C} \sum_{d=1}^{D} \Lambda_{cjj}\, V_{lj}\, Z_{cil} + 2 \sum_{c=1}^{C} \sum_{d,l=1}^{D} \sum_{m,n=1}^{D} U_{dm}\, \Lambda_{cmm}\, V_{nm} G_{cln}\, \Lambda_{cjj}\, V_{lj}\, \delta_{id} \tag{B.3c}$$

$$= -2 \sum_{c=1}^{C} \sum_{k,l=1}^{D} Z_{cil}\, V_{lk}\, \Lambda_{ckj}\, \delta_{kj} + 2 \sum_{c=1}^{C} \sum_{l,m,n=1}^{D} U_{im}\, \Lambda_{cmm}\, V_{nm} G_{cln}\, V_{lj}\, \Lambda_{cjj} \tag{B.3d}$$

$$= -2 \sum_{c} \mathbf{Z}_c \mathbf{V} \boldsymbol{\Lambda}_c + 2 \sum_{c} \mathbf{U} \boldsymbol{\Lambda}_c \mathbf{V}^\top \mathbf{G}_c \mathbf{V} \boldsymbol{\Lambda}_c \tag{B.3e}$$

B.1.3 Derivative w.r.t. V_{ij}

$$\frac{\partial}{\partial \boldsymbol{V}_{ij}} F(\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{V}) =$$

$$= -2 \sum_{c=1}^{C} \sum_{t \in c} \sum_{d=1}^{D} \left(x_{td} - (\boldsymbol{U}\, \boldsymbol{\Lambda}_c\, \boldsymbol{V}^\top \mu_{s_t})_d \right) \cdot U_{dj}\, \Lambda_{cjj}\, \mu_{s_t i} \tag{B.4a}$$

$$= -2 \sum_{c=1}^{C} \sum_{d=1}^{D} U_{dj}\, \Lambda_{cjj}\, Z_{cdi} + 2 \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{m,n=1}^{D} U_{dm}\, \Lambda_{cmm}\, V_{nm} G_{cnj}\, U_{dj} \Lambda_{cjj} \tag{B.4b}$$

$$= -2 \sum_{c=1}^{C} \sum_{d=1}^{D} Z_{cid}^\top U_{dj}\, \Lambda_{cjj} + 2 \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{m,n=1}^{D} G_{cin}\, V_{nm}\, \Lambda_{cmm} U_{md}^\top\, U_{dj} \Lambda_{cjj} \tag{B.4c}$$

$$= -2 \sum_{c} \boldsymbol{Z}_c^\top \mathbf{U} \boldsymbol{\Lambda}_c + 2 \sum_{c} \mathbf{G}_c \mathbf{V} \boldsymbol{\Lambda}_c \mathbf{U}^\top \mathbf{U} \boldsymbol{\Lambda}_c \tag{B.4d}$$

Appendix C

Corpora

Verbmobil II

Verbmobil was a major long-term German speech-to-speech translation research project funded by the German Ministry for Education, Science, Research and Technology (BMBF) and the industrial partners [Wahlster 00]. The consortium consisted of 31 universities, industrial companies and research institutes. The corpus of spontaneous dialogues in the domain of appointment scheduling and information desk [Burger & Weilhammer$^+$ 00]. Details of the Verbmobil corpus are given in Table C.1.

Table C.1: Statistics of the Verbmobil II training and test corpora.

Corpus	Training CD1-41	Test DEV99
Language	German	
Speaking Style	spontaneous	
Overall Duration [h]	61.5	1.6
Silence Fraction [%]	13	11
# Speakers	857	16
# Sentences	36 010	1 081
# Running Words	560 837	14 662
Class-Trigram LM Perplexity	-	62.0

Wall Street Journal – Spoke 3

The Wall Street Journal corpus was collected by the American National Institute of Standards and Technology for November '93 ARPA CSR II Hub and Spoke Benchmark Tests [Pallett & Fiscus$^+$ 94]. It consists of newspaper texts read by journalists from the Wall Street Journal. The Spoke 3 subset consists of non-native speakers

of American English (British, European, Asian dialects, etc.) and was used for the optional speaker adaptation evaluation (S3). Details of the corpus is given in Table C.2.

Table C.2: Statistics of WSJ1-Spoke3 training and test corpora.

Corpus	Training CSR-WSJ0	Test CSR-WSJ1 Spoke 3
Language	US English	US English (non-native)
Speaking Style	read	read
Overall Duration [h]	15.04	0.93
Silence Fraction [%]	18	26
# Speakers	83	40
# Sentences	7 138	400
# Running Words	129 435	6 611
Trigram LM Perplexity	-	71.9

North American Business News

The North American Business (NAB) corpus was collected by the American National Institute of Standards and Technology for November '94 ARPA CSR III Hub and Spoke Benchmark Tests [Pallett & Fiscus$^+$ 95]. It consists of newspaper texts from Reuters News Service, New York Times, Washington Post, Los Angeles Times and Wall Street Journal read by journalists.

Table C.3: Statistics of the NAB 20k training and test corpora.

Corpus	Training WSJ0+1	Test NAB DEV-94 H1
Language	US English	
Speaking Style	read	
Overall Duration [h]	81.4	0.8
Silence Fraction [%]	27	19
# Speakers	284	20
# Sentences	37 571	310
# Running Words	643 754	7 387
Trigram LM Perplexity	-	124.5

Aurora 4 – Noisy Wall Street Journal 5k

The Aurora 4 corpus is based on the WSJ0 corpus described above. Different noise samples at various signal-to-noise ratios (SNR) were added to simulate noisy data from the data recorded under clean studio conditions [Hirsch 02]. The noise was a collection of car, babble, restaurant, street, airport and train noise.

In this work, only the clean training data have been used for training, i.e. the original WSJ0 training data without added noise; the test results given in this work are averaged over all 14 official test sets.

Table C.4: Statistics of the Aurora 4 training and test corpora.

Corpus	Training CSR-WSJ0	Test CSR-WSJ0
Language	US English	
Speaking style	read	
Environment	studio	added noises
Overall Duration [h]	15.04	0.35
Silence Fraction [%]	18	26
# Speakers	83	8
# Sentences	7138	166
# Running Words	129 435	2 737
Trigram LM perplexity	–	62.3

Bibliography

[Acero 90] A. Acero: *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, Sept. 1990.

[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. 2, pp. 893–896, Toronto, Canada, May 1991.

[Afify & Siohan 00] M. Afify, O. Siohan: Constrained Maximum Likelihood Linear Regression for Speaker Adaptation. Proc. *Int. Conf. on Spoken Language Processing*, Vol. 3, pp. 861–864, Beijing, China, Oct. 2000.

[Ahadi & Woodland 97] S. Ahadi, P. Woodland: Combined Bayesian and Predictive Techniques for Rapid Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, Vol. 11, No. 3, pp. 187–206, July 1997.

[Alleva & Huang$^+$ 96] P. Alleva, X.D. Huang, M.Y. Hwang: Improvements on the Pronunciation Prefix Tree Search Organization. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 133–136, Atlanta, GA, USA, May 1996.

[Anastasakos & Balakrishnan 98] T. Anastasakos, S.V. Balakrishnan: The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers. Proc. *Int. Conf. on Spoken Language Processing*, Vol. 6, pp. 2303–2306, Sydney, NSW, Australia, Dec. 1998.

[Anastasakos & McDonough$^+$ 96] T. Anastasakos, H. McDonough, R. Schwartz, J. Makhoul: A Compact Model for Speaker-Adaptive Training. Proc. *Int. Conf. on Spoken Language Processing*, pp. 1137–1140, Philadelphia, PA, USA, Oct. 1996.

[Axelrod & Olsen 02] S. Axelrod, R.G.P. Olsen: Modeling with a Subspace Constraint on Inverse Covariance Matrices. Proc. *Int. Conf. on Spoken Language Processing*, pp. 2177–2180, Denver, CO, Sept. 2002.

[Bahl & Jelinek+ 83] L.R. Bahl, F. Jelinek, R.L. Mercer: A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Trans. 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, editor, *Speech Recognition*, pp. 512–542. Academic Press, New York, NY, USA, 1975.

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

[Baum 72] L.E. Baum: An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. In O. Shisha, editor, *Inequalities*, Vol. 3, pp. 1–8. Academic Press, New York, NY, 1972.

[Bayes 1763] T. Bayes: An Essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Society of London*, Vol. 53, pp. 370–418, 1763. Reprinted in *Biometrika*, vol. 45, no. 3/4, pp. 293–315, December 1958.

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

[Beulen & Ortmanns+ 99] K. Beulen, S. Ortmanns, C. Elting: Dynamic Programming Search Techniques for Across-Word Modeling in Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 609–612, Phoenix, AZ, USA, March 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*, pp. 692–695, Minneapolis, MN, USA, April 1993.

[Brown 87] P.F. Brown: *The Acoustic-Modeling Problem in Automatic Speech Recognition*. Ph.D. thesis, Carnegie Mellon University, May 1987.

[Burger & Weilhammer+ 00] S. Burger, K. Weilhammer, F. Schiel, H.G. Tillmann: *Verbmobil: Foundations of Speech-to-Speech Translation*, chapter "Verbmobil Data Collection and Annotation", pp. 537–549. Springer, Berlin, 2000.

[Chen & Eide+ 99] S.S. Chen, E.M. Eide, M.J.F. Gales, R.A. Gopinath, D. Kanevsky, P. Olsen: Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News. Proc. *DARPA Broadcast News Workshop*, pp. 89–93, Herndon, VA, USA, Feb. 1999.

[Chesta & Siohan+ 99] C. Chesta, O. Siohan, C.H. Lee: Maximum A Posteriori Linear Regression for Hidden Markov Model Adaptation. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 211–214, Budapest, Hungary, Sept. 1999.

[Chou 99] W. Chou: Maximum a Posterior Linear Regression with Elliptically Symmetric Matrix Variate Priors. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 1–4, Budapest, Hungary, Sept. 1999.

[Choukri & Chollet+ 86] K. Choukri, G. Chollet, Y. Grenier: Spectral Transformation Through Canonical Correlation Analysis for Speaker Adaptation in ASR. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 2659–2662, Tokyo, 1986.

[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. 3, pp. 1127–1130, Rhodes, Greece, Sept. 1997.

[Class & Kaltenmeier+ 90] F. Class, A. Kaltenmeier, P. Regel, K. Troller: Fast Speaker Adaptation for Speech Recognition Systems. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 133–136, Albuquerque,NM, April 1990.

[Cox 00] S. Cox: Speaker Normalization in the MFCC Domain. Proc. *Int. Conf. on Spoken Language Processing*, Vol. 2, 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 Trans. on Speech and Audio Processing*, Vol. 28, pp. 357–366, Aug. 1980.

[Demko & Moss+ 84] S. Demko, W. Moss, P. Smith: Decay rates for Inverse Band Matrices. *Mathematics of Computation*, Vol. 43, No. 168, pp. 491–499, Oct. 1984.

[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*, Vol. 39, No. 1, pp. 1–38, 1977.

[Digalakis & Berkowitz+ 99] V. Digalakis, S. Berkowitz, E. Bocchieri, C. Boulis, W. Byrne, H. Collier, A. Corduneanu, A. Kannan, S. Khudanpur, A. Sankar: Rapid Speech Recognizer Adaptation to New Speakers. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 765–768, Phoenix, AZ, USA, May 1999.

[Digalakis & Rtischev$^+$ 95] V. Digalakis, D. Rtischev, L. Neumeyer: Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures. *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 5, pp. 357–366, Sept. 1995.

[Ding & Zhu$^+$ 02] G.H. Ding, Y.F. Zhu, C. Li, B. Xu: Implementing Vocal Tract Length Normalization in the MLLR Framework. Proc. *Int. Conf. on Spoken Language Processing*, pp. 1389–1392, Denver, CO, Sept. 2002.

[Doddington 89] G. Doddington: Phonetically Sensitive Discriminants for Improved Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 556–559, Glasgow, UK, May 1989.

[Doddington & Przybocki$^+$ 00] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds: The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective. *Speech Communication*, Vol. 31, No. 2–3, pp. 225–254, June 2000.

[Doumpiotis & Tsakalidis$^+$ 04] V. Doumpiotis, S. Tsakalidis, W. Byrne: Discriminative Linear Transforms for Feature Normalization and Speaker Adaptation in HMM Estimation. Proc. *IEEE Trans. on Speech and Audio Processing*, 2004. To appear.

[Duda & Hart$^+$ 01] R.O. Duda, P.E. Hart, D.G. Stork: *Pattern Classification*. John Wiley & Sons, New York, NY, USA, 2001.

[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. 1, pp. 346–349, Atlanta, GA, May 1996.

[Emori & Shinoda 01] T. Emori, K. Shinoda: Rapid Vocal Tract Length Normalization using Maximum Likelihood Estimation. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 3, pp. 1649–1652, Aalborg, Denmark, Sept. 2001.

[Evermann & Chan$^+$ 04] G. Evermann, H. Chan, M. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, P. Woodland: Development of the 2003 CU-HTK Conversational Telephone Speech Transcription System. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 249–252, Montreal, Canada, May 2004.

[Fisher 36] R.A. Fisher: The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, Vol. 7, No. 179-188, 1936.

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

[Gales 96] M. Gales: The generation and use of regression class trees for MLLR adaptation. Technical Report CUED/F-INFENG/TR263, Cambridge University, 1996. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.

[Gales 97] M.J.F. Gales: Semi-Tied Full-Covariance Matrices for Hidden Markov Models. Technical Report CUED/F-INFENG/TR287, University of Cambridge, April 1997. (ftp://svr-ftp.eng.cam.ac.uk/pub/reports/gales_tr298.ps.gz).

[Gales 98] M.J.F. Gales: Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. *Computer Speech and Language*, Vol. 12, No. 2, pp. 75–98, April 1998.

[Gales 99] M.J.F. Gales: Semi-Tied Covariance Matrices for Hidden Markov Models. *IEEE Trans. on Speech and Audio Processing*, Vol. 7, No. 3, pp. 272–281, May 1999.

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

[Gales & Woodland 96] M.J.F. Gales, P.C. Woodland: Mean and Variance Adaptation within the MLLR framework. *Computer Speech and Language*, Vol. 10, No. 4, pp. 249–264, Oct. 1996.

[Gao & Ramabhadran$^+$ 00] Y. Gao, B. Ramabhadran, M. Picheny: New Adaptation Techniques For Large Vocabulary Continuous Speech Recognition. Proc. *ISCA Automatic Speech Recognition Workshop*, pp. 107–111, Paris, France, Sept. 2000.

[Gauvain & Lee 94] J. Gauvain, C.H. Lee: Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chainns. *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291–298, April 1994.

[Generet & Ney$^+$ 95] M. Generet, H. Ney, F. Wessel: Extensions to Absolute Discounting for Language Modeling. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 2, pp. 1245–1248, Madrid, Spain, Sept. 1995.

[Gopinath 98] R.A. Gopinath: Maximum Likelihood Modeling with Gaussian Distributions for Classification. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 661–664, Seattle, WA, May 1998.

[Gunawardana & Byrne 01] A. Gunawardana, W. Byrne: Discounted likelihood linear regression for rapid speaker adaptation. Computer, Speech and Language. *Computer Speech and Language*, Vol. 15, No. 1, pp. 15–38, 2001.

[Haeb-Umbach 01] R. Haeb-Umbach: Automatic Generation of Phonetic Regression Class Trees for MLLR Adaptation. *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 3, March 2001.

[Haeb-Umbach & Ney 92] R. Haeb-Umbach, H. Ney: Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 13–16, San Francisco, CA, USA, March 1992.

[Haeb-Umbach & Ney 94] R. Haeb-Umbach, H. Ney: Improvements in Beam Search for 10000-Word Continuous-Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 353–356, April 1994.

[Hermansky 90] H. Hermansky: Perceptual Linear Predictive (PLP) Analysis of Speech. *Journ. Acoustic Soc. America*, Vol. 87, No. 4, pp. 1738–1752, June 1990.

[Hilger 04] F. Hilger: *Quantile Based Histogram Equalization for Noise Robust Speech Recognition*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, 2004.

[Hilger & Ney 01] F. Hilger, H. Ney: Quantile Based Histogram Equalization for Noise Robust Speech Recognition. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 2, pp. 1135–1138, Aalborg, Denmark, Sept. 2001.

[Hilger & Ney$^+$ 03] F. Hilger, H. Ney, O. Siohan, F.K. Soong: Combining Neighboring Filter Channels to Improve Quantile Base d Histogram Equalization. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. I, pp. 640–643, Hong Kong, China, April 2003.

[Hirsch 02] H.G. Hirsch: Experimental Framework for the Performance Evaluation of Speech Recognition Front–ends on a Large Vocabulary Task, Version 2.0, AU/417/02. Technical report, ETSI STQ-Aurora DSR Working Group, Oct. 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, USA, 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.

[Hunt 79] M. Hunt: A Statistical Approach to Metrics for Word and Syllable Recognition. *J. Acoust. Soc. Am.*, Vol. 66(S1), No. S35(A), Nov. 1979.

[Jaschul 82] J. Jaschul: Speaker Adaptation by a Linear Transformation with Optimised Parameters. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1657–1670, Paris, France, May 1982.

[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. *Proceedings of the IEEE*, Vol. 64, No. 10, pp. 532–556, April 1976.

[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*, pp. 1531–1534, Istanbul, Turkey, June 2000.

[Katz 87] S.M. Katz: Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Trans. on Speech and Audio Processing*, Vol. 35, pp. 400–401, March 1987.

[Kumar & Andreou 98] N. Kumar, A.G. Andreou: Heteroscedastic Discriminant Analysis And Reduced Rank HMMs For Improved Speech Recognition. *Speech Communication*, Vol. 26, No. 4, pp. 283–297, Dec. 1998.

[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. 1, pp. 353–356, Atlanta, GA, May 1996.

[Leggetter 95] C.J. Leggetter: *Improved Acoustic Modelling for HMMs Using Linear Transformations*. Ph.D. thesis, University of Cambridge, 1995.

[Leggetter & Woodland 95a] 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, No. 2, pp. 171–185, April 1995.

[Leggetter & Woodland 95b] C. Leggetter, P. Woodland: Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. Proc. *ARPA Spoken Language Technology Workshop*, pp. 104–109, 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 Technical Journal*, Vol. 62, No. 4, pp. 1035–1074, April 1983.

[Lowerre 76] B. Lowerre: *A Comparative Performance Analysis of Speech Understanding Systems*. Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, PA, USA, 1976.

[McDonough 98] J.W. McDonough: Speaker Normalization With All-Pass Transforms. Technical Report 28, Center for Language Speech Processing, The Johns Hopkins University, Baltimore, MD, Sept. 1998. (http://www.clsp.jhu.edu/people/jmcd/postscript/all-pass.ps).

[McDonough 00] J.W. McDonough: *Speaker Compensation With All-Pass Transforms*. Ph.D. thesis, Johns Hopkins University, Baltimore, MD, 2000.

[McDonough & Zavaliagkos$^+$ 96] J.W. McDonough, G. Zavaliagkos, H. Gish: An Approach to Speaker Adaptation based on Analytic Functions. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 721–724, Atlanta, GA, May 1996.

[Molau 02] S. Molau: Private Communication. RWTH Aachen, 2002.

[Molau 03] S. Molau: *Normalization in the Acoustic Feature Space for Improved Speech Recognition*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, 2003.

[Molau & Kanthak$^+$ 00] S. Molau, S. Kanthak, H. Ney: Efficient Vocal Tract Normalization in Automatic Speech Recognition. Proc. *Proc. Elektronische Sprachsignalverarbeitung ESSV*, pp. 209–216, Cottbus, Germany, Sept. 2000.

[Molau & Keysers$^+$ 02] S. Molau, D. Keysers, H. Ney: Matching Training and Test Data Distributions for Robust Speech Recognition. *Speech Communication*, Vol. 41, No. 4, pp. 579–601, 2002.

[Molau & Pitz$^+$ 01] 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. 1, pp. 73–76, Salt Lake City, UT, June 2001.

[Nene & Nayar 96] S.A. Nene, S.K. Nayar: Closest Point Search in High Dimensions. Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 859–865, San Francisco, CA, USA, June 1996.

[Neumeyer & Sankar$^+$ 95] L. Neumeyer, A. Sankar, V. Digalakis: A Comparative Study of Speaker Adaptation Techniques. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 2, pp. 1127–1130, Madrid, Sept. 1995.

[Ney 84] H. Ney: The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 32, No. 2, pp. 263–271, April 1984.

[Ney 90] H. Ney: Acoustic Modeling of Phoneme Units for Continuous Speech Recognition. In L. Torres, E. Masgrau, M.A. Lagunas, editors, *Signal Processing*

V: Theories and Applications, Fifth European Signal Processing Conference, pp. 65–72. Elsevier Science Publishers B. V., Barcelona, Spain, 1990.

[Ney & Aubert 94] H. Ney, X. Aubert: A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition. Proc. *Int. Conf. on Spoken Language Processing*, Vol. 3, pp. 1355–1358, Yokohama, Japan, Sept. 1994.

[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 & 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. 1, pp. 9–12, San Francisco, CA, USA, March 1992.

[Ney & Martin$^+$ 97] H. Ney, S.C. Martin, F. Wessel: Statistical Language Modeling using Leaving-One-Out. In S. Young, G. Bloothooft, editors, *Corpus Based Methods in Language and Speech Processing*, pp. 1–26. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[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*, pp. 833–836, Dallas, TX, USA, April 1987.

[Ney & Oerder 93] H. Ney, M. Oerder: Word Graphs: An Efficient Interface Between Continuous Speech Recognition and Language Understanding. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 119–122, Minneapolis, MN, USA, April 1993.

[Nguyen & Gelin$^+$ 99] P. Nguyen, P. Gelin, J.C. Junqua, J.T. Chien: N-Best Based Supervised and Unsupervised Adaptation for Native and Non-Native Speakers in Cars. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 173–176, Phoenix, AZ, USA, March 1999.

[Nguyen & Rigazio$^+$ 03] P. Nguyen, L. Rigazio, J.C. Junqua: Large Corpus Experiments for Broadcast News Recognition. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, pp. 1837–1840, Geneva, Switzerland, Sept. 2003.

[Nocerino & Rabiner$^+$ 85] F.K. Nocerino, L.R. Rabiner, D.H. Klatt: Comparative Study of Several Distortion Measures for Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 25–28, Atlanta, GA, April 1985.

[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, USA, March 1994.

[Olsen & Gopinath 02] P.A. Olsen, R.A. Gopinath: Modeling Inverse Covariance Matrices by Basis Expansion. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 945–948, Orlando, FL, May 2002.

[Ortmanns 98] S. Ortmanns: *Effiziente Suchverfahren zur Erkennung kontinuierlich gesprochener Sprache*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, Nov. 1998.

[Ortmanns & Ney 95] S. Ortmanns, H. Ney: An Experimental Study of the Search Space for 20000-Word Speech Recognition. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 2, pp. 901–904, Madrid, Spain, Sept. 1995.

[Ortmanns & Ney+ 96a] S. Ortmanns, H. Ney, A. Eiden: Language-Model Look-Ahead for Large Vocabulary Speech Recognition. Proc. *Int. Conf. on Spoken Language Processing*, Vol. 4, pp. 2095–2098, Philadelphia, PA, USA, Oct. 1996.

[Ortmanns & Ney+ 96b] S. Ortmanns, H. Ney, A. Eiden, N. Coenen: Look-Ahead Techniques for Improved Beam Search. Proc. *CRIM-FORWISS Workshop*, pp. 10–22, Montreal, Canada, Oct. 1996.

[Ortmanns & Ney+ 97a] S. Ortmanns, H. Ney, X. Aubert: A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition. *Computer Speech and Language*, Vol. 11, No. 1, pp. 43–72, Jan. 1997.

[Ortmanns & Ney+ 97b] S. Ortmanns, H. Ney, T. Firzlaff: Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 139–142, Rhodes, Greece, Sept. 1997.

[Padmanabhan & Saon+ 00] M. Padmanabhan, G. Saon, G. Zweig: Lattice-Based Unsupervised MLLR for Speaker Adaptation. Proc. *ISCA Automatic Speech Recognition Workshop*, pp. 128–132, Paris, Sept. 2000.

[Pallett & Fiscus+ 94] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, M.A. Przybocki: 1993 Benchmark Tests for the ARPA spoken language program. Proc. *ARPA Human Language Technology Workshop*, pp. 49–54, Plainsboro, NJ, USA, March 1994.

[Pallett & Fiscus+ 95] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, M.A. Przybocki: 1994 Benchmark Test for the ARPA spoken language program. Proc. *ARPA Human Language Technology Workshop*, pp. 5–36, Austin, TX, USA, Jan. 1995.

[Paul 91] D.B. Paul: Algorithms for an Optimal A^* Search and Linearizing the Search in the Stack Decoder. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 693–696, Toronto, Canada, May 1991.

[Press & Teukolsky$^+$ 02] W. Press, S. Teukolsky, W. Vetterling, B. Flannery: *Numerical Recipes in C++*. Cambridge University Press, 2002.

[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. 2, pp. 1046–1050, Munich, Germany, April 1997.

[Rabiner 89] L.R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286, Feb. 1989.

[Rabiner & Juang 86] L. Rabiner, B.H. Juang: An Introduction to Hidden Markov Models. *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 1, pp. 4–16, 1986.

[Rabiner & Schafer 78] L.R. Rabiner, R.W. Schafer: *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[Ramasubramansian & Paliwal 92] V. Ramasubramansian, K.K. Paliwal: Fast k-dimensional Tree Algorithms for Nearest Neighbor Search with Application to Vector Quantization Encoding. *IEEE Trans. on Speech and Audio Processing*, Vol. 40, No. 3, pp. 518–528, March 1992.

[Sakoe 79] H. Sakoe: Two-Level DP-Matching - A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 27, pp. 588–595, Dec. 1979.

[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. 1, pp. 121–124, Adelaide, Australia, April 1995.

[Sankar & Lee 96] A. Sankar, C.H. Lee: A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, pp. 190–202, May 1996.

[Saon & Zweig$^+$ 03] G. Saon, G. Zweig, B.K.L. Mangu, U. Chaudhari: An Architecture for Rapid Decoding of Large Vocabulary Conversational Speech. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, pp. 1977–1980, Geneva, Switzerland, Sept. 2003.

[Schlüter 02] R. Schlüter: Private Communication. RWTH Aachen, 2002.

[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. 1, pp. 701–704, Toronto, Canada, May 1991.

[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*, pp. 81–84, Albuquerque, NM, USA, April 1990.

[Schwartz & Colthurst$^+$ 04] R. Schwartz, T. Colthurst, N. Duta, H. Gish, R. Iyer, C.L. Kao, D. Liu, O. Kimball, J. Ma, J. Makhoul, S. Matsoukas, L. Nguyen, M. Noamany, R. Prasad, B. Xiang, D.X. Xu, J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen: Speech Recognition in Multiple Languages and Domains: The 2003 BBN/LIMSI Ears System. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 753–756, Montreal, Canada, May 2004.

[Shinoda & Lee 97] K. Shinoda, C.H. Lee: Structural MAP Speaker Adaptation Using Hierarchical Priors. Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 381–388, Santa Barbara, CA, USA, Dec. 1997.

[Sixtus 03] A. Sixtus: *Across-Word Phoneme Models for Large Vocabulary Continuous Speech Recognition.* Ph.D. thesis, RWTH Aachen, Jan. 2003.

[Sixtus & Ney 02] A. Sixtus, H. Ney: From Within-Word Model Search to Across-Word Model Search in Large Vocabulary Continuous Speech Recognition. *Computer Speech and Language*, Vol. 16, No. 2, pp. 245–271, May 2002.

[Soltau & Schaaf$^+$ 01] H. Soltau, T. Schaaf, F. Metze, A. Waibel: The Isl Evaluation System For Verbmobil-II. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, June 2001.

[Steinbiss & Tran$^+$ 94] V. Steinbiss, B.H. Tran, H. Ney: Improvements in Beam Search. Proc. *Int. Conf. on Spoken Language Processing*, Vol. 4, pp. 2143–2146, Yokohama, Japan, Sept. 1994.

[Uebel & Woodland 99] L.F. Uebel, P.C. Woodland: An Investigation into Vocal Tract Length Normalisation. Proc. *ISCA Europ. Conf. on Speech Communication and Technology*, Vol. 6, pp. 2527–2530, Budapest, Hungary, Sept. 1999.

[Uebel & Woodland 01] L.F. Uebel, P.C. Woodland: Speaker Adaptation Using Lattice-Based MLLR. Proc. *ISCA ITR-Workshop on Adaptation Methods in Speech Recognition*, pp. 57–60, Sophia Antipolis, France, Aug. 2001.

[Vintsyuk 71] T.K. Vintsyuk: Elementwise Recognition of Continuous Speech Composed of Words from a Specified Dictionary. *Kibernetika*, 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, pp. 260–269, 1967.

[Wahlster 00] W. Wahlster, editor: *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag, Berlin, Germany, 2000.

[Wakita 77] H. Wakita: Normalization of Vowels by Vocal Tract Length and its Application to Vowel Identification. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-25, No. 2, pp. 183–192, April 1977.

[Wallhoff & Willet$^+$ 00] F. Wallhoff, D. Willet, G. Rigoll: Frame-Discriminative and Confidence-Driven Adaption for LVCSR. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1835–1838, Istanbul, Turkey, June 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. 1, pp. 339–341, Atlanta, GA, May 1996.

[Welling 99] L. Welling: *Merkmalsextraktion in Spracherkennungssystemen für großen Wortschatz*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, Jan. 1999.

[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. 2, pp. 761–764, Phoenix, AZ, April 1999.

[Welling & Ney$^+$ 02] L. Welling, H. Ney, S. Kanthak: Speaker Adaptive Modeling by Vocal Tract Normalization. *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 6, pp. 415–426, Sept. 2002.

[Wessel 02] F. Wessel: *Word Posterior Probabilities for Large Vocabulary Continuous Speech Recognition*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, Aug. 2002.

[Wessel & Macherey$^+$ 98] F. Wessel, K. Macherey, R. Schlüter: Using Word Probabilities as Confidence Measures. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 225–228, Seattle, WA, USA, May 1998.

[Wessel & Ortmanns$^+$ 97] F. Wessel, S. Ortmanns, H. Ney: Implementation of Word Based Statistical Language Models. Proc. *Spoken Queries in European Languages (SQEL) Workshop on Multi-Lingual Information Retrieval Dialogs*, pp. 55–59, Pilsen, Czech Republic, April 1997.

[Wessel & Schlüter$^+$ 00] F. Wessel, R. Schlüter, H. Ney: Using Posterior Probabilities for Improved Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1587–1590, Istanbul, Turkey, June 2000.

[Wessel & Schlüter$^+$ 01] F. Wessel, R. Schlüter, K. Macherey, H. Ney: Confidence Measures for Large Vocabulary Continuous Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 3, pp. 288–298, March 2001.

[Woodland 01] P.C. Woodland: Speaker Adaptation for Continuous Density HMMs: A Review. Proc. *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, pp. 11–19, Sofia Antipolis, France, Aug. 2001.

[Young 92] S.J. Young: The General Use of Tying in Phoneme Based HMM Recognizers. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 569–572, San Francisco, CA, USA, March 1992.

[Young 93] S.J. Young. *HTK: Hidden Markov Model Toolkit V1.4. User Manual.* Cambridge, UK, Feb. 1993.

[Zeppenfeld & Finke$^+$ 97] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, A. Waibel: Recognition of Conversational Telephone Speech Using the Janus Speech Engine. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1815–1818, Munich, April 1997.

[Zolnay 03] A. Zolnay: Private Communication. RWTH Aachen, 2003.

[Zolnay 04] A. Zolnay: Private Communication. RWTH Aachen, 2004.

Lebenslauf – Curriculum Vitae

Angaben zur Person

Name	Michael Pitz
Geburtsdatum	21.08.1970
Geburtsort	Aachen
Nationalität	deutsch
Familienstand	ledig

Schulbildung

1977 - 1981	Gemeinschaftsgrundschule Hermannstraße, Stolberg
1981 - 1990	Goethe-Gymnasium, Stolberg

Studium

1990 - 1996	Studium der Physik an der RWTH Aachen
1996 - 1998	Promotionsstudium Physik an der RWTH Aachen
1998 - 2004	Promotionsstudium Informatik an der RWTH Aachen

Berufserfahrung

1996 - 1998	Wissenschaftlicher Angestellter am II. Physikalischen Institut der RWTH Aachen
1998 - 2004	Wissenschaftlicher Angestellter am Lehrstuhl für Informatik VI der RWTH Aachen
seit 2004	Entwicklungsspezialist bei der BMW Forschung und Technik GmbH