Acoustic Feature Combination for Speech Recognition

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

vorgelegt von

Diplom–Ingenieur András Zolnay

aus

Budapest, Ungarn

Berichter: Universitätsprofessor Dr.–Ing. Hermann Ney
Universitätsprofessor Dr.–Ing. Reinhold Haeb-Umbach

Tag der mündlichen Prüfung: 14. August 2006

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.

To my parents, Ildikó and Gábor

Acknowledgments

First of all I would like to thank my supervisor, Prof. Dr.-Ing. Hermann Ney, head of the Chair of Human Language Technology and Pattern Recognition, Lehrstuhl für Informatik VI, at the RWTH Aachen-University, for his support and his interest. He introduced me to speech recognition in 2000 when I started my studies as a PhD student and he has since then given me the opportunity and the freedom to pursue my ideas.

I would also like to thank my second supervisor Prof. Dr.-Ing. Reinhold Häb-Umbach, head of the Department of Communications Engineering at the University of Paderborn, for his interest in this thesis and his valuable advice.

I am very grateful to Dr. rer.-nat. Ralf Schlüter for his support in the field of signal analysis and acoustic modeling. His supportive coaching helped me to make decisions and to define my long-term research goals. I am also very much indebted to my colleagues Max Bisani and Stephan Kanthak for their advises in research and software technology. Their revolutionary ideas and delight in programming escorted me through the past five years. I am enriched by the experience of developing the new RWTH research software SPRINT in cooperation with Max and Stephan. I would also like to thank Georg Heigold for his hard work in designing and implementing the Discriminative Model Combination toolkit. His contribution was essential for completing the goals of this thesis.

It should not be left unmentioned that I would not have enjoyed my last year without our two guest researchers. Prof. Umesh, from the Indian Institute of Technology, Kanpur, India, supported me with his experience and interest in feature extraction and especially in Vocal Tract Length Normalization. Daniil Kocharov, from Department of Phonetics, Saint-Petersburg State University, brought new ideas into my research on acoustic feature extraction. His knowledge on phonetics was essential to broaden the circle of acoustic features considered in this thesis. Most of all though, both of them were great colleagues and are still good friends.

I also have to thank our network and computer support team for their excellent support with our continuously improving computing equipment. In no particular order these include Oliver Bender, Thomas Deselaers, Stephan Kanthak, Mirko Kohns, and Stephan Koltermann. All the other people at the Lehrstuhl für Informatik VI are also deserving of my thanks for the discussions that we had, and for making work more fun.

Finally, I would like to thank my family for their understanding and encouragements.

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under the post graduate program "Software für Kommunikationssysteme" and by the European Commission under the project TC-Star (FP6-506738).

Abstract

In this thesis, the use of multiple acoustic features of the speech signal is considered for speech recognition. The goals of this thesis are twofold: on the one hand, new acoustic features are developed, on the other hand, feature combination methods are investigated in order to find an effective integration of the newly developed features into state-of-the-art speech recognition systems.

The most commonly used feature extraction methods are the Mel Frequency Cepstrum Coefficients (MFCC), Perceptual Linear Prediction (PLP), and variations of these techniques. These methods are mainly based on the models of the human auditory system. A detailed review of the implementation of these features is presented in this thesis. There have also been attempts at using articulatory motivated acoustic features for speech recognition which are motivated by models of the human speech production system. This thesis focuses partially on the development of new articulatory motivated acoustic features. The voicing information is one of the most commonly used articulatory features. Three voicing extraction methods are presented in this work followed by a systematic comparison. Besides the analysis of the voicing feature, the novel spectrum derivative feature is introduced which aims to capture the differences between magnitude spectra produced by obstruent and sonant consonants.

The articulatory motivated features are tested in combinations with state-of-the-art acoustic features based on auditory models mainly. The features are combined both directly using Linear Discriminant Analysis (LDA) as well as indirectly on model level using Discriminative Model Combination (DMC). Both methods have already been used successfully in automatic speech recognition systems. In this work, a comparative study is presented which describes and analyzes the application of these methods to feature combination. Robustness issues of the LDA based method are addressed which are induced by increasing the amount of acoustic features coefficients. An application of DMC to feature combination is introduced based on the splitting of the acoustic model into separate scalable knowledge sources. After the analysis of the individual methods, a comparison is carried out on the basis of the underlying acoustic emission models.

Experimental results are presented for small- and large-vocabulary tasks. The results show that the accuracy of automatic speech recognition systems can be significantly improved by the combination of auditory and articulatory motivated features. The combination of the Vocal Tract Length Normalized MFCC and articulatory motivated features demonstrates that additional articulatory information can even improve the performance of speaker adapted systems. The word error rate is reduced from 1.8% to 1.5% on the *SieTill*, a German digit string recognition task. Consistent improvements in word error rate have been obtained on two large-vocabulary corpora. The word error rate is reduced from 19.1% to 18.2% on the *VerbMobil II*, a German large vocabulary conversational speech task, and from 14.1% to 13.5% on the European Parliament Plenary Sessions task.

Zusammenfassung

In dieser Arbeit wird die Verwendung mehrerer akustischer Merkmale des Sprachsignals für die Erkennung kontinuierlich gesprochener Sprache untersucht. Die Arbeit hat zwei Zielsetzungen:: Einerseits werden neue akustische Merkmale entworfen, andererseits werden Merkmalskombinationstechiken untersucht, um die neuen Merkmale in moderne Spracherkennungssysteme effektiv zu integrieren.

Die am meisten verbreiteten akustischen Merkmale sind Mel Frequency Cepstrum Coefficients (MFCC), Perceptual Linear Prediction (PLP) und Variationen dieser Merkmale. Diese Verfahren beruhen hauptsächlich auf Modellen des menschlichen Gehörsystems. Die vorliegende Arbeit enthält einen ausführlichen Überblick über die Implementierung dieser akustischen Merkmale. Neben Merkmale motiviert durch Modelle des Gehörsystems sind zahlreiche artikulatorische Merkmalsextraktionsverfahren untersucht worden, die auf Modelle der menschlichen Sprachproduktion beruhen. In dieser Arbeit werden neue artikulatorische Merkmale des Sprachsignals vorgestellt. Das Merkmal Stimmhaftigkeit ist ein sehr verbreitetes artikulatorisches Merkmal. Drei Extraktionsverfahren des Merkmals Stimmhaftigkeit werden in dieser Arbeit vorgestellt und systematisch verglichen. Neben dem Merkmal Stimmhaftigkeit wird das neue Merkmal differenziertes Spektrum vorgestellt. Dieses Merkmal fokussiert auf Unterschiede in den Amplitudenspektren von Verschlußkonsonanten und stimmhaften Konsonanten.

Die artikulatorischen Merkmale werden in Kombination mit gängigen akustischen Merkmalen getestet, die hauptsächlich auf Modellen des menschlichen Gehörsystems beruhen. Zwei Kombinationsverfahren werden im Rahmen dieser Arbeit getestet: Linear Discriminant Analysis (LDA) und Discriminative Model Combination (DMC). Beide Verfahren sind schon in automatischen Spracherkennungssystemen erfolgreich eingesetzt worden. In dieser Arbeit wird eine vergleichende Untersuchung durchgeführt, die die Anwendung dieser Methoden in der akustischen Merkmalskombination in den Vordergrund stellt. Die Robustheit der LDA-basierten Methode wird bei steigender Anzahl von akustischen Merkmalskoeffizienten analysiert. Das DMC-Verfahren wird auf die Merkmalskombination angewendet, wobei das akustische Modell in einzelne skalierbare Informationsquellen zerlegt wird. Nach der Analyse der Methoden werden die zu Grunde liegenden Emissionsmodelle verglichen.

Der Einfluss von unterschiedlichen Merkmalen auf die Fehlerrate wird anhand zahlreicher Experimente mit Korpora kleinen und großen Wortschatzes studiert. Erkennungsergebnisse demonstrieren, dass die Fehlerrate automatischer Spracherkennungssysteme mit der Kombination von gehörbasierten und artikulatorischen Merkmalen signifikant verbessert werden kann. Die Kombination von dem vokaltraktlängennormierten MFCC und artikulatorischen Merkmalen zeigt, daß zusätzliche artikulatorische Merkmale sogar die Erkennungsleistung sprecherabhängiger Systeme verbessern können. Die Wortfehlerrate wird auf dem deutschen Ziffernerkennungskorpus *SieTill* von 1.8% auf 1.5% gesenkt. Konsistente Verbesserungen in der Wortfehlerrate werden auch auf zwei Korpora mit

großem Wortschatz erzielt. Die Wortfehlerrate wird auf dem deutschen Spontansprachkorpus *VerbMobil II* von 19.1% auf 18.2% und auf dem EPPS-Korpus (Aufnahmen aus Plenarsitzungen des Europäischen Parlaments) von 14.1% auf 13.5% gesenkt.

Contents

1 Introduction **1**

1.1 Statistical Speech Recognition 1
1.2 Signal Analysis 2
1.3 Acoustic Modeling 3
1.4 Language Modeling 6
1.5 Search 7

2 State of the Art **11**

2.1 Acoustic Feature Extraction 11
2.2 Acoustic Feature Combination 13
2.3 Summary 16

3 Scientific Goals **17**

4 Signal Analysis **21**

4.1 Mel Frequency Cepstral Coefficients 21
4.2 Perceptual Linear Predictive Analysis 24
4.3 PLP Derived from Mel Scale Filter Bank 27
4.4 Vocal Tract Length Normalization 29
4.4.1 Implementation of Frequency Warping 30
4.4.2 Training of Warping Factors 36
4.4.3 Training of Acoustic Model 37
4.4.4 Recognition Using VTLN 38
4.5 Voicing Feature 39
4.5.1 Harmonic Product Spectrum 39
4.5.2 Autocorrelation 42
4.5.3 Average Magnitude Difference 45

4.5.4 Comparison of Voicing Measures . . . 47
4.5.4.1 Experimental Results . . . 48
4.5.4.2 Analytical Relation of AC and HPS . . . 48
4.5.4.3 Analytical Relation of AC and AMD . . . 50
4.5.4.4 Summary . . . 50
4.5.5 Optimization of Autocorrelation Method . . . 51
4.5.5.1 Frame Length . . . 51
4.5.5.2 Estimation of Autocorrelation Function . . . 51
4.5.6 Recognition Results . . . 54
4.5.7 Summary . . . 56
4.6 Spectrum Derivative Feature . . . 57
4.6.1 Extraction Algorithm . . . 57
4.6.2 Optimization of Spectrum Derivative Measure . . . 58
4.6.3 Results and Analysis . . . 61
4.6.4 Summary . . . 62

5 LDA Based Feature Combination 63
5.1 Theory of Linear Discriminant Analysis . . . 64
5.2 Feature Combination Algorithm . . . 67
5.3 Experimental Setup . . . 68
5.4 Recognition Results . . . 68
5.5 Analysis of Results . . . 70
5.5.1 Combination with White Noise Components . . . 71
5.5.2 Sensitivity of Eigenvalues and Vectors . . . 73
5.5.2.1 Definition of Sensitivity of Eigenvalues and Vectors . . . 73
5.5.2.2 Experimental Results . . . 76
5.5.3 Comparison of Confusion Matrices . . . 77
5.5.4 Summary . . . 79

6 DMC Based Feature Combination 81
6.1 Feature Combination Algorithm . . . 81
6.2 Training of DMC Based Feature Combination . . . 83
6.2.1 Training of Emission Feature Function . . . 84
6.2.1.1 Initialization . . . 84
6.2.1.2 Training Algorithm . . . 86

6.2.1.3 Iteration of Training Algorithm . . . 87
6.2.2 Training of Model Weights . . . 88
6.2.2.1 Training Algorithm . . . 88
6.2.2.2 Calculation of Log-Linear Feature Values . . . 90
6.2.2.3 Initialization of Model Weights . . . 90
6.3 Integration of Log-Linear Model . . . 91
6.4 Recognition Tests . . . 94
6.5 Summary . . . 96

7 Comparison of LDA and DMC Based Combinations 99
7.1 Theoretical Comparison . . . 99
7.1.1 State Tying Structure . . . 101
7.1.2 Amount of Combination Parameters . . . 101
7.2 Comparison of Recognition Results . . . 102
7.3 Summary . . . 104

8 Scientific Contributions 105

9 Outlook 109

A Corpora and Recognition Systems 111
A.1 SieTill . . . 111
A.2 VerbMobil II . . . 112
A.3 European Parliament Plenary Sessions . . . 112
A.4 Summary . . . 113

B Detailed Calculations 115
B.1 Derivation of MWE criterion . . . 115

C Symbols and Acronyms 119
C.1 Mathematical Symbols . . . 119
C.2 Acronyms . . . 123

Bibliography 125

List of Tables

4.1 Word error rates for baseline acoustic features. 30

4.2 Word error rate on the *SieTill* and *VM II* corpora obtained by the combination of the MFCC feature with one of the voicing measures: harmonic product spectrum (HPS), autocorrelation (AC), or average magnitude difference (AMD). (Note that a within-word acoustic model has been used here for the *VM II* corpus.) 48

4.3 Word error rate on the *VM II* corpus obtained by increasing the window length W of the voicing measure. The tests have been carried out by combining the MFCC feature with two different voicedness measures: autocorrelation, average magnitude difference. (Note that the within-word acoustic model has been used here.) 52

4.4 Word error rate on the *VM II* corpus obtained by using auto- and cross-correlation while calculating the voicing measures. The tests have been carried out by combining the MFCC feature with two different voicedness measures: autocorrelation, average magnitude difference. (Note that the within-word acoustic model has been used here.) 54

4.5 Word error rates obtained by the LDA based combination of baseline features (MFCC, VTLN, MF-PLP, PLP) and the voicing feature (V). . . . 55

4.6 Overlap of histograms shown on Fig. 4.24. 55

4.7 Word error rates obtained by the LDA based combination of the MFCC and the spectrum derivative (SD) features. The window length of the spectrum derivative measure has been increased from 15ms to 90ms. Experiments have been performed on the *SieTill* corpus. 59

4.8 Word error rates obtained by increasing the cut-off frequency of the low-pass filter. The spectrum derivative measure is tested in the LDA based combination with the MFCC feature. 60

4.9 Word error rates obtained by increasing number of spectrum derivative measures. The spectrum derivative measures are tested in the LDA based combination with the MFCC and the voicing (V) features. Note when adding the ith order spectrum derivative, all the lower order $(1, 2, ..., i\text{-}1)$ ones are used as well. 61

4.10 Word error rates obtained by the LDAbased combination of baseline features (MFCC, VTLN) and the spectrum derivative feature (SD). 62

5.1 Consistent improvements in word error rates obtained by the LDA based combination of increasing number of acoustic features. 69

5.2 Degradation in word error rate obtained by the LDA based combination of baseline features MFCC, MF-PLP, and PLP. 70

5.3 Word error rate obtained by the LDA based combination of MFCC features and increasing number of randomly generated coefficients. 72

5.4 Average reciprocal condition numbers and asymptotic error bounds of eigenvalues and eigenvectors on the *EPPS* corpus obtained by LDA based feature combination tests yielding improvements and degradations in word error rate. 77

5.5 Average reciprocal condition numbers and asymptotic error bounds of eigenvalues and eigenvectors on the *EPPS* corpus obtained by using increasing LDA window lengths. 77

5.6 Difference confusion matrix generated by subtracting the confusion matrix obtained by using the LDA based combination of the MFCC and voicing features from the one obtained by using solely the MFCC feature. 78

6.1 Word error rate obtained by applying discriminative model combination (DMC) to single and LDA-combined features. 94

6.2 Model weights obtained by applying discriminative model combination to single and LDA-combined features. (All model weights are divided by the language model weight.) 95

7.1 Number of CART classes or CART index tuples obtained when combining the same acoustic features with using the LDA and DMC based feature combination methods. 101

7.2 Word error rates obtained by applying Discriminative Model Combination (DMC) based and Linear Discriminant Analysis (LDA) to the same set of features. 103

A.1 Settings of the RWTH recognition system for the *SieTill*, *VM II*, and *EPPS* corpora. 113

List of Figures

1.1 Architecture of a statistical automatic speech recognition system 2

1.2 6-state hidden Markov model in Bakis topology 5

4.1 Block diagram of Mel Frequency Cepstral Coefficients. 21

4.2 Linear filter bank in Mel-frequency domain and Mel warped filter bank in original frequency domain. 23

4.3 Block diagram of Perceptual Linear Predictive Analysis. 24

4.4 Trapezoidal filter bank in Bark-frequency domain and Bark-warped filter bank in original frequency domain. For sake of comparability, the warped frequency $\tilde{\omega}$ has been normalized by $2\pi \cdot 600$Bark-Hz similar to Eq. (4.8) . 25

4.5 Mel $g_{Mel}(\omega)$ Eq. (4.5) and Bark $g_{Bark}(\omega)$ Eq. (4.8) warping functions. For sake of comparability, the Bark warping function has been scaled by an a such that $g_{Mel}(2\pi \cdot 8\text{kHz}) = a \cdot g_{Bark}(2\pi \cdot 8\text{kHz})$. 28

4.6 Block diagram of MF-PLP feature. 28

4.7 Comparison of dynamic range compression functions. The PLP feature uses the intensity loudness law $\sqrt[3]{x}$ whereas the MFCC feature uses the logarithm $\ln(x)$. 29

4.8 Piece-wise linear warping function with warping factor α. 31

4.9 Block diagram of vocal tract length normalized MFCC feature (VTLN) using separate warping nodes. 32

4.10 Block diagram of vocal tract length normalized MFCC feature (VTLN) using integrated warping. 35

4.11 Shrunken magnitude spectra and the harmonic product spectrum of a 40ms long time frame from the middle of an utterance of the sound /a/. 40

4.12 Harmonic product spectrum of two 40ms long time frames from the middle (left) and from the end (right) of an utterance of the sound /a/. 41

4.13 Block diagram of the harmonic product spectrum based voicing measure. . 42

4.14 Histograms of the measure $v_{HPS}^{[t]}$ estimated on the *VM II* corpus. Left: voiced plosive sound /g/ and its unvoiced counterpart /k/. Right: voiced fricative sound /v/ and its unvoiced counterpart /f/. 42

4.15 Windowed time frame of a voiced speech signal. 43

4.16 Unbiased estimate of the autocorrelation of two 40ms long time frames from the middle (left) and from the end (right) of an utterance of the sound /a/. 44

4.17 Block diagram of the autocorrelation based voicing measure. 44

4.18 Histograms of the measure $v_{AC}^{[t]}$ estimated on the *VM II* corpus. Left: voiced plosive sound /g/ and its unvoiced counterpart /k/. Right: voiced fricative sound /v/ and its unvoiced counterpart /f/. 45

4.19 Average magnitude difference of two 40ms long time frames from the middle (left) and from the end (right) of an utterance of the sound /a/. 45

4.20 Block diagram of the average magnitude spectrum based voicing measure. . 47

4.21 Histograms of measure $v_{AMD}^{[t]}$ estimated on the *VM II* corpus. For easier comparability with Fig. 4.14 and Fig. 4.18, values of $1 - v_{AMD}^{[t]}$ have been used instead of the original ones. Left: voiced plosive sound /g/ and its unvoiced counterpart /k/. Right: voiced fricative sound /v/ and its unvoiced counterpart /f/. 47

4.22 Windowing scheme for the cross-correlation based voicing measure. 52

4.23 Block diagram of the cross-correlation based voicing measure. 53

4.24 Histograms of LDA output coefficients belonging to the three largest eigenvalues. Histograms have been estimated on the *VM II* corpus for the voiced fricative sound /v/ and its unvoiced counterpart /f/. Left: first three output coefficients (y_0, y_1, y_2) of LDA applied solely to the MFCC feature. Right: first three output coefficients $(\tilde{y}_0, \tilde{y}_1, \tilde{y}_2)$ of the LDA based combination of the MFCC and voicing features. 56

4.25 Block diagram of spectrum derivative measure. 57

4.26 Histograms of spectrum derivative measure $s^{[t]}$ for the sonant consonant /v/ and the obstruent consonant /s/ estimated on the *VM II* corpus. . . . 61

6.1 Histogram of a-posteriori probability Eq. (6.12) of the best sentence in the N-best lists for $\eta = 3$ on the *EPPS* development corpus. 91

Chapter 1

Introduction

1.1 Statistical Speech Recognition

In recent years, the statistical approach to speech recognition has prevailed over other approaches, as summarized in [Pitz 2005]. 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]:

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

The two basic stochastic models are defined in Eq. (1.1) which 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 1990]. 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).

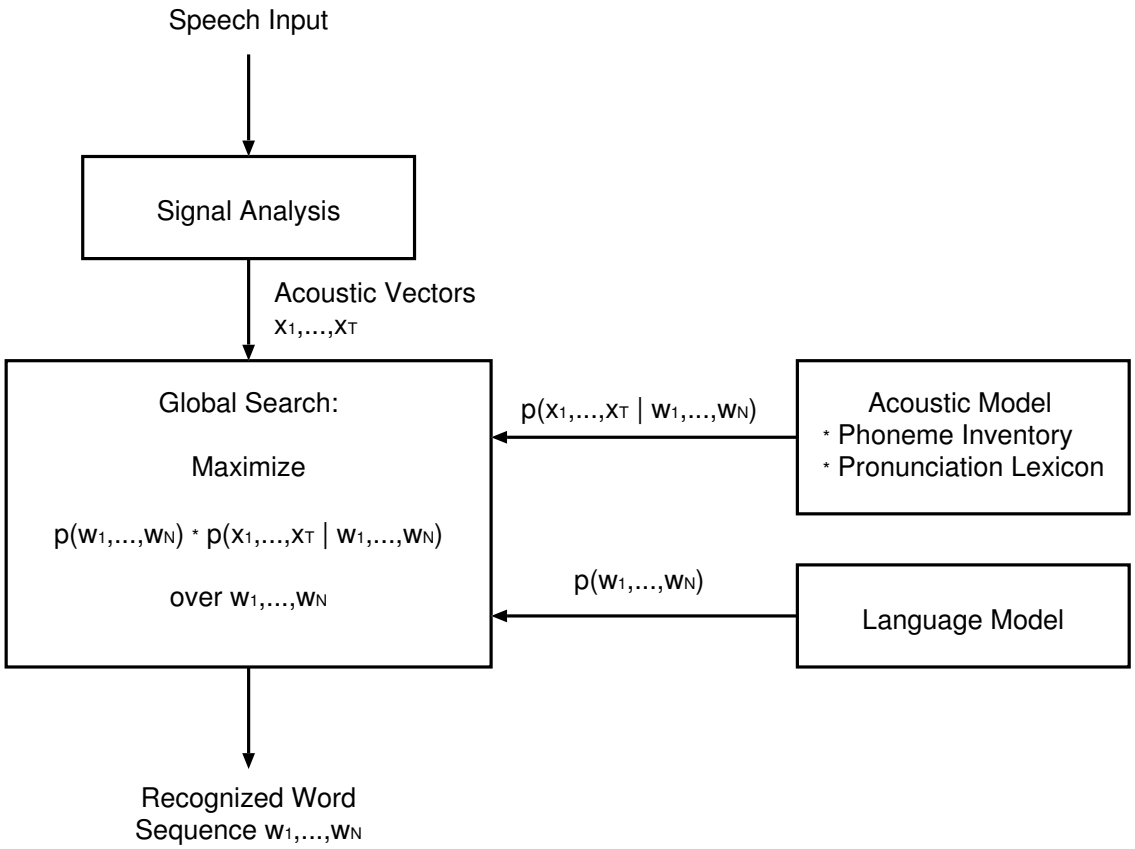


Figure 1.1: Basic architecture of a statistical automatic speech recognition system. [Ney 1990].

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 parametrization 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, background noise, speaker identity, 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 and Schafer 1978], usually a Fourier analysis. Three procedures for further processing and smoothing are widely used: *Mel frequency Cepstral coefficients* (MFCC) [Davis and Mermelstein 1980], *Perceptual Linear Prediction* (PLP) [Hermansky 1990], and the combination of these two methods first published in [Woodland et al. 1997]. In this work, a detailed review of these methods is presented in Chapter 4. Beside features derived form the short-term power spectrum, several alternative acoustic features have been developed in the recent years. A comprehensive overview of acoustic feature extraction is presented in Chapter 2 as well. The aforementioned features are motivated by the models of the human auditory system. In this work, a study on the voicing and on the novel spectrum derivative feature is presented in Chapter 4. Both features are based on models of the human speech production system, also called as articulatory models. The articulatory motivated features are applied to recognition tasks in combination with the state-of-the-art features. Combination of recognition systems have been tested at different levels. An overview of the state-of-the-art acoustic feature combination is presented in Chapter 2.

A commonly used method to include dynamic information is augmenting the original feature vector with the first and second derivatives yielding a high dimensional vector. A more general approach is based on the *Linear Discriminant Analysis* (LDA) [Fisher 1936, Duda et al. 2001]. 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 (seven to eleven vectors, depending on the specific task). This high dimensional feature space is reduced using LDA to a dimension of typically ≈ 40.

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 et al. 2000], which means that there is still plenty of information of the given speaker contained in those features. Several methods have been developed to cope with the speaker dependence of the 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. In [Pitz 2005], a comprehensive comparison of these methods is presented along with a unified view of speaker dependent transformations.

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 1975, Rabiner 1989]. A 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 et al. 2001]. 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 1990]

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

The Eqs. (1.2) and (1.3) can be solved efficiently using the *forward-backward* algorithm [Baum 1972, Rabiner and Juang 1986] or *dynamic programming* [Bellman 1957, Viterbi 1967, Ney 1984].

An example of a HMM for a part of the word "seven" is given in Figure 1.2. The topology used in this work has been introduced by Bakis [Bakis 1976]: 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. Chapter 4) 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 2003]. 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.

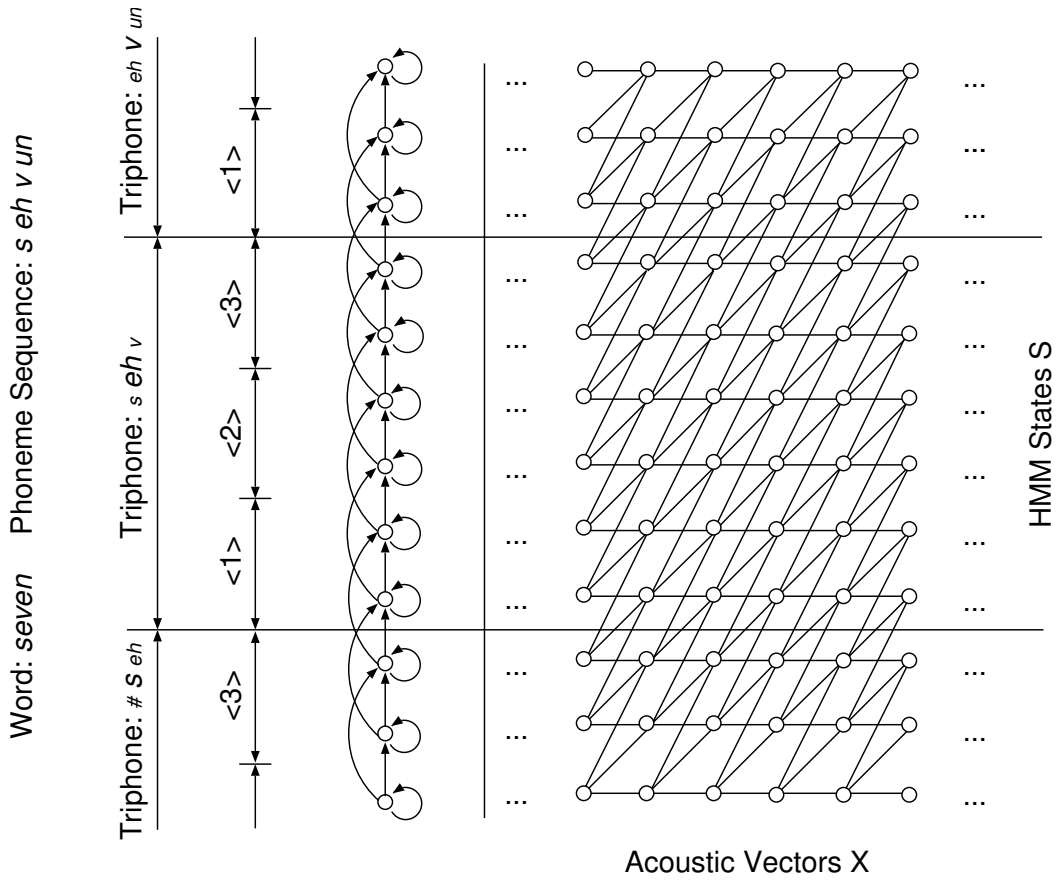


Figure 1.2: 6-state hidden Markov model in Bakis topology for the triphone ${}_{s}eh_{v}$ in the word "seven". The HMM segments are denoted by <1>, <2>, and <3>.

The emission probabilities $p(x_t|s_t; w_1^N)$ of a HMM can be modeled by discrete probabilities [Jelinek 1976], semi-continuous probabilities [Huang and Jack 1989] or as continuous probability distributions [Levinson et al. 1983]. 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 et al. 1999]. 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 (i.e. all states share the same variance) 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 et al. 1977].

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 have no or too few observations for a reliable parameter estimation. Therefore several states are tied together yielding *generalized* n-phone models [Young 1992]. 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 et al. 1999]. 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 and Lee 1991, Odell et al. 1994], 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 2003].

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 et al. 1983] have become widely accepted. The 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 1987, Ney et al. 1994, Generet et al. 1995, Ney et al. 1997]. 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 et al. 1994]. Details of the language model implementation in the RWTH system can be found in [Wessel et al. 1997].

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)

$$\begin{aligned}[w_1^N]_{\text{opt}} &= \underset{w_1^N}{\text{argmax}}\, p(w_1^N|x_1^T) \\ &= \underset{w_1^N}{\text{argmax}} \left\{p(w_1^N) \cdot p(x_1^N|w_1^N)\right\} . \end{aligned} \tag{1.6}$$

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

$$\begin{aligned}[w_1^N]_{\text{opt}} &= \underset{w_1^N}{\text{argmax}} \left\{ \left[\prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1})\right] \cdot \left[\sum_{s_1^T : w_1^N} \prod_{t=1}^{T} p(x_t|s_t; w_1^N) \cdot p(s_t|s_{t-1}; w_1^N)\right] \right\} \\ &\cong \underset{w_1^N}{\text{argmax}} \left\{ \left[\prod_{n=1}^{N} p(w_n|w_{n-m+1}^{n-1})\right] \cdot \left[\max_{s_1^T : w_1^N} \prod_{t=1}^{T} p(x_t|s_t; w_1^N) \cdot p(s_t|s_{t-1}; w_1^N)\right] \right\} . \end{aligned} \tag{1.7}$$

In the second step the maximum (Viterbi) approximation has been applied to the HMM, which significantly reduces the complexity of the optimization problem. Eq. (1.7) can be solved efficiently using *dynamic programming* [Bellman 1957]. Dynamic programming exploits the mathematical structure and divides the problem into sub-instances. 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 1969, Paul 1991]. 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 1971, Baker 1975, Sakoe 1979, Ney 1984]. 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. (1.7) 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. (1.7) 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 1976, Ney et al. 1987, Ortmanns and Ney 1995]. 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 reduce further the computational complexity of the Viterbi or beam-search:

- *Lexical prefix tree* [Ney et al. 1992]: 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 et al. 1993, Odell et al. 1994, Alleva et al. 1996, Ortmanns et al. 1996a]: 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 et al. 1992, Häb-Umbach and Ney 1994, Ortmanns et al. 1996b].

- *Fast likelihood computation:* a considerable amount of computation time of a 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 and Paliwal 1992, Fritsch 1997], quantization of the feature vectors [Bocchieri 1993, Ortmanns et al. 1997b] or by partitioning the feature space [Nene and Nayar 1996]. Details of the fast likelihood computation used in the RWTH system can be found in [Ortmanns 1998]. Additionally, utilizing the Single Instruction Multiple Data (SIMD) technique of modern CPUs for a parallelized likelihood computation significantly reduces the computation time [Kanthak et al. 2000].

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 and Chow 1990, Schwartz and Austin 1991]. A word graph is a directed acyclic graph whose arcs contain the word labels of word sequence alternatives [Ney and Oerder 1993, Ney and Aubert 1994, Ortmanns et al. 1997a]. 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 et al. 2004, Evermann et al. 2004]. 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 et al. 2003]. Subsequently, 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 by using these adapted acoustic models.

Chapter 2

State of the Art

Acoustic feature extraction has a long history in speech recognition research. Combination of different recognition systems became subject of increasing interest in the recent years. This chapter gives an overview of the most important approaches related to this thesis. In Section 2.1, acoustic feature extraction techniques are reviewed followed by the summary of state-of-the-art acoustic feature combination techniques in Section 2.2.

2.1 Acoustic Feature Extraction

In this section, a comprehensive review of state-of-the-art auditory and articulatory motivated feature extraction techniques is presented for speech recognition. Short summaries of publications are provided on those acoustic features which are used in this investigation, and which are closely related to the topic of this thesis.

The most widespread acoustic features are mainly based on models of the human auditory system. Some constraining properties of the human hearing such as nonlinear (Bark or Mel) frequency scale, spectral amplitude compression, decreasing sensitivity of hearing at lower frequencies, and large spectral integration are already integrated in the state-of-the-art acoustic features for automatic speech recognition. One of the most often used acoustic features, the Mel Frequency Cepstrum Coefficients (MFCC) feature was first introduced in [Davis and Mermelstein 1980]. The Perceptual Linear Predictive (PLP) feature introduced in [Hermansky 1990] is based on ideas similar to the MFCCs. Nevertheless, there are major differences in data flow and in recognition performance as well. The third fairly widespread auditory based feature has been derived from the two aforementioned ones, as described in [Woodland et al. 1997]. The MF-PLP feature uses a Mel scale triangular filter bank embedded into the data flow of the PLP feature. The *Minimum Variance Distortionless Response* (MVDR) was first investigated in [Murthi and Rao 1997] to improve the estimation of the power spectrum of a speech signal. Recently, the aforementioned three methods have been tested by using MVDR at different stages of the extraction algorithms. A systematic comparison of power spectrum estimates have been presented in [Wölfel et al. 2003]. In [Dharanipragada and Rao 2001], improvements in word error rate (WER) of $\approx 30\%$ have

been observed when applying MVDR to a noisy corpus.

Besides methods processing the short-term magnitude spectrum, new acoustic features have been proposed recently which focus on the short-term phase spectrum [Paliwal 2003]. On a small-vocabulary task, significant reduction in WER has been presented in [Schlüter and Ney 2001] by using the combination of the MFCCs and a set of smoothed phase features. Applying discriminative training, the additional features have resulted in an improvement in WER of 23% relative to using the MFCCs alone. Acoustic features derived from the group delay function have already been researched in different speech applications [Zhu and Paliwal 2004, Murthy and Gadde 2003, Hegde et al. 2005]. In [Hegde et al. 2004], improvements in WER of up to 18% have been reported when combining a modified group delay function based feature with the MFCCs on two Indian broadcast news corpora.

Applications of articulatory models have already been intensively studied in speech recognition systems. The first related studies go back to rule based speech recognition. The integration of articulatory information has been researched at different stages of a speech recognition system. The methods vary from front-end feature extraction algorithms to complex synthesis models replacing the state-of-the-art acoustic modeling.

The voicing feature is one of the most commonly used articulatory features. In rule-based speech recognition systems, voiced-unvoiced detection has been used as one of the acoustical features. In [Atal and Rabiner 1976], a voiced-unvoiced-silence detection algorithm is proposed using statistical approaches. A voicing measure instead of a voiced-unvoiced decision is described in [Thomson and Chengalvarayan 1998]. The authors presented results obtained by combining an autocorrelation based voicing measure, a voicing jitter measure and LPC-derived liftered cepstral coefficients. Using the concatenated features along with their first and second derivatives, a large relative improvement in WER of 40% has been achieved by applying discriminative training. Different voicing measure extraction methods have been compared in [Zolnay et al. 2003]. Recognition tests have been carried out by using a Harmonic Product Spectrum, an autocorrelation and an Average Magnitude Difference based voicing measures along with the MFCC feature. The experimental comparison has shown that the three extraction methods do not differ significantly. In this thesis, the voicing feature is applied to different feature combination setups. After a detailed description of the aforementioned three extraction methods, a theoretical comparison is presented in Section 4.5.4 along with experimental results. In [Graciarena et al. 2004], the entropy of high order cepstrum coefficients is used to extract voicing information. Recognition tests have been performed of the NIST Hub-5 dev2001 and eval2002 databases. An average improvement in WER of 2% has been obtained realtive to the best single feature system when the entropy based voicing feature has been combined with an autocorrelation based voicing and with the MFCC features. These experimental results comply with the results presented in this thesis.

In [Welling and Ney 1996], one of the first recognition systems was presented which use formant frequencies as acoustic features. Features derived form the position of formants have replaced the MFCCs. Although the formant based features performed unexpectedly well, they could not outperform the baseline features on an English

digit recognition task. Using only 5 formant based coefficients has yielded a WER of 1.7% whereas using 16 MFCC coefficients has resulted in a WER of 0.6%. Feature combination experiments have not been presented within the scope of this publication. In [Padmanabhan 2000, Weber et al. 2001, Stuttle and Gales 2002], formant frequencies have been used in combination with the MFCC feature. One of the first experimental results has been published in [Holmes et al. 1997]. Significant improvements in WER have been obtained on a connected-digit recognition task when adding the formant based features. The WER has been improved by 33% relative to using 5 cepstrum coefficients and the energy. Nevertheless, the results have to be considered as preliminary since an acoustic model with single densities has been used to perform the recognition tests.

A novel acoustic feature has been proposed recently in [Kocharov et al. 2005] to distinguish obstruent from sonant consonants. The *spectrum derivative* feature captures the differences in the magnitude spectrum of these two phoneme classes. Sonants manifest in peaky formant-like magnitude spectra whereas obstruents typically produce flat and noisy ones. Hence, a measure summarizing the intensity of changes of the magnitude spectrum over the frequency axis can help to differentiate these two phoneme classes. The spectrum derivative feature has been tested in combination with the MFCC feature yielding improvements in WER of up to 11% on small- and of up to 3% on large-vocabulary tasks relative to using the MFCCs alone.

Articulatory information can be successfully utilized to improve the MFCC feature. In [Gu and Rose 2001], magnitude spectrum values falling on harmonics have been emphasized yielding large improvement in WER on a noisy Mandarin isolated digit string recognition task. The idea of scaling the frequency axis of the speech signal to account for gender specific variation of the vocal tract has been proposed first in [Wakita 1977]. Meanwhile, Vocal Tract Length Normalization (VTLN) became a standard method in the speech recognition community [Acero 1990, Lee and Rose 1996, Wegmann et al. 1996]. The implementation of VTLN in the RWTH speech recognition system is shortly reviewed in Section 4.4. The application of VTLN improves the WER by up to 10% relative to using the speaker independent MFCC feature.

In the following, a few examples are mentioned for the application of articulatory information to acoustic modeling. In [Eide 2001], the ratio of "absent" and "present" probabilities of articulatory classes has been used as acoustic feature. The "absent" and "present" distributions have been trained over the MFCC feature. The probability ratios of a few articulatory classes (e.g. voice, strident, vocalic, continuant) have been concatenated with MFCC coefficients. The combined features have resulted in a large improvement in WER of 33% on a corpus containing city and street names. An explicit hidden dynamic model of the coarticulation process has been first introduced in [Picone et al. 1999], using a synthesis model in a rescoring scheme. During rescoring, MFCC feature vectors generated by articulatory based hidden dynamic models have been compared with the observed ones. Similar approaches have been presented in [Richards and Bridle 1999, Y. Gao and Xiang 2000].

In Chapter 4, a subset of the aforementioned acoustic features is presented which has been used in different feature combination experiments. The reviews include the MFCC, PLP, and MF-PLP features. Subsequently, the implementation of the Vocal Tract Length

Normalization is introduced. Two articulatory motivated features are studied in detail. The implementation of the voicing and the spectrum derivative measures are presented along with recognition results obtained on different corpora.

2.2 Acoustic Feature Combination

The goal of acoustic feature combination is to find an optimal way of utilizing the mutually complementary classification information of different features. The combination can be carried out directly on the level of feature vectors, via acoustic probabilities, or at the level of recognition results. In the following, a review of those state-of-the-art feature combination techniques is presented which are closely related to the methods investigated in this work.

Combination of acoustic features can be performed directly on the level of feature vectors. In [Holmes et al. 1997, Thomson and Chengalvarayan 1998, Eide 2001], the baseline cepstral feature vectors have been directly concatenated with additional articulatory motivated features. In most of the systems, the concatenated feature vectors are also augmented by their first and second order derivatives. A more general approach is based on the Linear Discriminant Analysis (LDA). In [Fisher 1936, Duda et al. 2001], LDA has been introduced for feature space reduction. In [Häb-Umbach and Ney 1992], LDA has been used successfully to find an optimal linear combination of successive vectors of a single-feature stream. The linear combination of successive feature vectors can be seen as the generalization of the first and second order derivatives. Beside the combination of successive feature vectors, LDA can be applied to feature combination tasks as well. The combination of different cepstral features has been tested by using the LDA in [Häb-Umbach and Loog 1999]. Here, LDA has not only found an optimal linear combination of successive vectors of a single feature stream but three different cepstral features have also been combined by using a single linear transformation. Time-synchronous vectors of the three feature streams have been concatenated first followed by the concatenation of successive concatenated vectors. In every time frame, the resulting large feature vector has been projected onto a ≈ 40 dimensional subspace by using the LDA matrix. Nevertheless, the combination of three cepstral features has not yielded any significant improvements in WER compared to using the MFCCs alone. Significant reduction in WER has been obtained by using LDA to combine acoustic features with fairly different information content. In [Schlüter and Ney 2001], the LDA based combination of the MFCCs and a set of phase spectrum based features has resulted in an improvement in WER of 23% relative to using solely the MFCCs. On the same small-vocabulary corpus, the LDA based combination of the MFCC and the voicing features has yielded an improvement in WER of 11% [Zolnay et al. 2002]. Despite the successful application of LDA to feature extraction, there are still open questions left on the robustness of the algorithm. In [Welling 1999], experiments have been presented with increasing LDA window length i.e. with increasing number of successive concatenated feature vectors. In contrast to the expectations, a clear optimum has been found at ≈ 11 concatenated feature vectors. Theoretically, increasing the LDA window length (e.g. increasing the information content of the LDA) should not lead to degradation

in WER. In [Katz et al. 2002], among other robustness issues the same open problem has been investigated using artificially generated data. In Chapter 5, the results in [Katz et al. 2002] are reviewed along with new findings on the robustness of LDA. An analysis of the numerical stability of LDA is presented by means of the perturbation theory. Furthermore, new questions are handled which have emerged from the application of LDA to feature combination.

Combination of acoustic features can also be carried out at the level of acoustic probabilities. In this case, acoustic models trained on different features produce probabilities which are combined by either their product or sum. In hybrid HMM and artificial neural network based recognition systems, the feature combination is mainly carried out by averaging the logarithm of the a-posteriori probabilities produced by the different acoustic neural networks. In [Kirchhoff 1998, Janin et al. 1999, Ellis and Bilmes 2000], different auditory and articulatory based acoustic features have been combined yielding significant improvements in WER. In [Shire 2001], scaling of the combined a-posteriori probabilities is introduced, however, without the automatic training of the model scales. A more general approach for the combination of knowledge sources and respectively for feature combination is the log-linear model. The basic idea is to modify the modeling of the posterior probability $P(W|X)$ in the Bayes' decision rule by applying the log-linear model to the a-posteriori probability:

$$P(W|X) = \frac{e^{\sum_i \lambda_i g_i(W,X)}}{\sum_{W'} e^{\sum_i \lambda_i g_i(W',X)}}, \tag{2.1}$$

where g_i is a so called *feature function* which is an arbitrary function of the word sequence W and the feature vector sequence X, and λ_i is the corresponding log-linear weight. In standard speech recognition systems, the feature functions are the negative logarithm of emission, transition, pronunciation, and language model probabilities. The log-linear distribution allows the use of not only a single instance of each of these knowledge sources but a combination of several instances. The log-linear model combination has already been applied to different problems in speech recognition. In [Cook and Robinson 1996], the adaptation of the boosting algorithm is described to the acoustic training. The acoustic models trained on the different sub-sets of the training corpus are combined in a log-linear manner. Experiments using GMMs trained for different articulatory classes are described in [Metze and Waibel 2002]. The GMMs have been trained on MFCC feature for $\approx$ 10 different articulatory classes. The standard acoustic probabilities and the probabilities provided by the articulatory GMMs have been combined using log-linear combination. On a large-vocabulary task, the additional articulatory based acoustic models have yielded an improvement in WER of 13% relative to using a single acoustic model. In [Tolba et al. 2002], acoustic features have been combined by the means of log-linear modeling. The MFCC feature combined with auditory acoustic cues and a main spectral peak feature has yielded a significant reduction in WER on the *TIMIT* corpus. Word error minimizing training of log-linear model weights for speech recognition models was proposed in [Beyerlein 1997]. One of the the first successful application of the Discriminative Model Combination (DMC) was the automitic estimation of the language model scale. Five acoustic and language models have been combined by using DMC in [Beyerlein 1998]. The combination of within-word and across-word acoustic models and

bigram, trigram, and fourgram language models has resulted in an improvement in WER of 5% compared to the best pairwise combinations. One of the first applications of DMC to acoustic feature combination was published in [Häb-Umbach and Loog 1999]. The combination of three cepstral features (MFCC, MF-PLP, and LPC-smoothed MFCC) has been tested on the Hub-4 corpus. An improvement in WER of 3% has been obtained relative to using the MFCC feature alone. In Chapter 6, beside recognition test using the combination of auditory and articulatory motivated features, a detailed description of DMC is presented along with its application to feature combination.

System and respectively feature combination can also be carried out at the level of recognition results. For the sake of completeness, a simple voting is mentioned here which operates on the recognized word sequences. The method known as Recognition Output Voting Error Reduction (ROVER) was first published in [Fiscus 1997]. However the comparison presented in [Beyerlein 2000, Häb-Umbach and Loog 1999] has shown that ROVER performs similar or is inferior to DMC when applied to the same large-vocabulary task.

2.3 Summary

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

- Most of the automatic speech recognition systems use either the MFCC or the PLP feature. Although many alternatives have been proposed, no single feature system has performed better consistently on several large-vocabulary corpora.
- Articulatory motivated features can improve the performance of speech recognition systems however mostly in combination with one of the baseline features.
- Beside the simple concatenation of feature vectors, the LDA based feature combination is one of the simplest feature combination methods operating directly on feature vectors. It has been successfully applied to single and multi-feature systems. However, there are open problems on the robustness of method.
- System combination is one of the emerging techniques which have already been applied successfully to speech recognition. Oracle experiments have shown that the combination of significantly different systems can improve the WER considerably.
- DMC has been recently developed for the automatic combination of recognition systems. Besides the first attempts, its application to feature combination has not been investigated yet.
- Although the LDA and DMC based feature combination methods operate at different levels of a speech recognition system, they produce emission models with similar structures. However, no theoretical or experimental comparison has been published yet.

Chapter 3

Scientific Goals

As discussed in the previous chapter, numerous auditory and articulatory motivated acoustic feature have already been developed which focus on different cues in the speech signal. One of the promising ways of utilizing mutually complementary classification information in different acoustic features is to apply an appropriate combination method. In this chapter, conclusions are drawn from the state-of-the-art of acoustic feature extraction and combination. The goal of this thesis is to investigate the combination of several acoustic features. A review of the extraction of several state-of-the-art auditory motivated acoustic features is provided followed by detailed descriptions of the extraction of the voicing and the novel spectrum derivative features. Two feature combination methods are studied: the direct combination of feature coefficients by using Linear Discriminant Analysis (LDA) and the combination of features via their acoustic emission probabilities using Discriminative Model Combination (DMC). Both methods have already been used in automatic speech recognition systems however the interconnection has not been addressed yet. In particular, the following aspects are considered in detail:

Voicing measure

Former investigations showed that the incorporation of the voicing information into speech recognition systems can improve the word error rate. In this thesis, three methods are studied to find an appropriate extraction of the voicing information. The common goal of the methods is to provide a measure of voicing instead of a voiced-unvoiced decision. To this end, the following basic functions are applied: Harmonic Product Spectrum, autocorrelation, and Average Magnitude Difference. These functions have already been applied to numerous speech applications. They express the similarity between a speech segment and its shifted copy. However, the post-processing of the similarity functions is not obvious. In order to produce a bounded measure of voicing, the peak structure of the similarity functions has to be analyzed and a normalization has to be applied to account for varying frame energies. In this work, the three voicing extraction methods are first compared on theoretical level followed by experiments performed in combination with different state-of-the-art acoustic features.

Spectrum derivative measure

The novel *spectrum derivative measure* has been developed in cooperation with Daniil Kocharov to distinguish obstruent from sonant consonants. The feature captures the differences in the magnitude spectrum of these two phoneme classes. In this thesis, the feature and its extraction are described and analyzed in details. Similar to the voicing feature, the appropriate integration of the feature is not obvious. Among other aspects, the normalization of the spectrum derivative measure is optimized with respect to the WER. The analysis is carried out by means of histograms estimated on obstruent-sonant phoneme pairs and experiments are performed to verify that the spectrum derivative measure can contribute to better recognition results on small- and large-vocabulary corpora.

Linear Discriminant Analysis

In former publications, *Linear Discriminant Analysis* (LDA) was successfully applied to single- and multi-feature speech recognition systems. However, there are open questions on the robustness of the method. Former publications have addressed the robustness of LDA against increasing dimension of the input vectors, see [Welling 1999] [Katz et al. 2002]. The authors have mainly studied the problem from the point of view of statistical parameter estimation and have drawn attention to the sparseness of the estimated scatter matrices. In this work, the problem of increasing dimensionality is discussed by means of the perturbation theory applied to the underlying eigenproblem. Furthermore, a broad review of the application of LDA to acoustic feature combination is presented. Pitfalls of the LDA based feature combination are revealed and explanations are provided by making use of the theory of the perturbation theory. Experiments performed on large-vocabulary corpora are analyzed not only in terms of word error rate but also in terms of sensitivity of the eigenvalue and eigenvector estimates.

Discriminative Model Combination

In earlier publications, *Discriminative Model Combination* (DMC) was mainly used to combine different acoustic and language models. There were also attempts at applying DMC to acoustic feature combination. However, a comprehensive study pointing out advantages and disadvantages has not been published yet. In this work, a detailed description is presented which introduces an application of DMC to feature combination. The basic idea of splitting the acoustic model is elaborated resulting in separate scalable knowledge sources for instance the emission, transition, and pronunciation models. The weights of the combined models are trained by optimizing the *Minimum Word Error* criterion. Beside the derivation of the optimization procedure, details of the initialization of the discriminative training are provided which turn out to be indispensable for a successful application of DMC. Experimental results are presented using the DMC based combination of auditory and articulatory motivated features. The experiments are not only analyzed in term of the WER but an insight is also provided into the resulting weights of the combined knowledge sources.

Comparison of LDA and DMC Based Combination Methods

Although the LDA based and the DMC based feature combination techniques operate at different levels of speech recognition systems, the resulting combined emission models share a common structure. In spite of a few empirical results, a comprehensive comparison of feature combination techniques has not been published yet. In this thesis, a theoretical comparison of the LDA and the DMC based methods is presented which considers the similarities and differences of the underlying emission models. The emission models are compared in terms of the triphone state tying structure and the number of parameters involved in the combination. The experimental comparison analyzes recognition results obtained by both methods using the same set of features. Advantageous and disadvantages setups are pointed out and explanations are provided.

The remainder of this thesis is organized as follows: In Chapter 4, a review is presented on the implementation of the MFCC, PLP, and MF-PLP features followed by a summary of the RWTH implementation of Vocal Tract Length Normalization. Detailed descriptions of the voicing and the spectrum derivative measures are presented here as well along with recognition results on different corpora. The LDA based feature combination method is summarized in Chapter 5. Beside feature combination experiments, robustness issues are also discussed by means of the perturbation theory. Chapter 6 describes the application of DMC to acoustic feature combination. A theoretical comparison of the two methods is carried out in Chapter 7 followed by the evaluation of comparative recognition test. This thesis will be concluded by a summary of the scientific contributions in Chapter 8 and an outlook in Chapter 9. Details on the speech corpora and recognition settings used in the experimental evaluations as well as details on some more complex calculations will be given in the Appendix.

Chapter 4

Signal Analysis

In this chapter, feature extraction methods are presented which have been used in different feature combination experiments. First, the implementation of the Mel Frequency Cepstrum Coefficients (MFCC) is described in Section 4.1, followed by the Perceptual Linear Predictive (PLP) feature and its alternative the MF-PLP feature in Section 4.2 and 4.3, respectively. Subsequently, in Section 4.4, the Vocal Tract Length Normalization is shortly reviewed. Finally, two articulatory motivated features are presented: Three extraction methods of the voicing feature are compared in Section 4.5. In Section 4.6, a recently developed new acoustic feature, the spectrum derivative feature is presented.

4.1 Mel Frequency Cepstral Coefficients

The data flow of the Mel Frequency Cepstral Coefficients (MFCC) feature extraction is depicted on Fig. 4.1. In the first step, a preemphasis of the sampled speech signal is

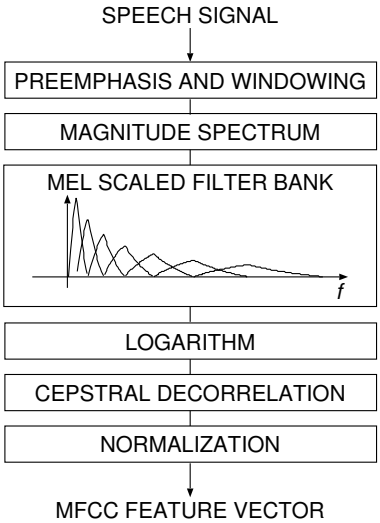


Figure 4.1: Block diagram of Mel Frequency Cepstral Coefficients.

performed. The preemphasized samples $d[n]$ are obtained from the original samples $s[n]$ by differencing:

$$d[n] = s[n] - s[n-1]. \tag{4.1}$$

Every 10ms, a Hamming window is applied to preemphasized 25ms long speech segments. The Hamming window is defined as follows:

$$h[t] = 0.54 - 0.46 \cdot \cos\left(2\pi \frac{t}{W-1}\right) \quad : \quad t = 0, \ldots, W-1 \tag{4.2}$$

where W denotes the discrete length of the window $W = 25\text{ms} \cdot f_s$ and f_s is the sample rate. The short-term spectrum is computed by using the Fast Fourier Transform (FFT) along with an appropriate zero padding. The minimum amount of inserted zeros is determined automatically. First, the number of FFT points is found by looking for the smallest integer N_{FFT} such that

$$2^{N_{FFT}} >= W : N_{FFT} \in \mathbb{N} \tag{4.3}$$

This inequality has a straightforward solution:

$$N_{FFT} = \left\lceil \frac{\log W}{\log 2} \right\rceil. \tag{4.4}$$

Hence, the number of zeros can now be determined: $N_{FFT} - W$. Next, Mel warping is applied to the magnitude spectrum. The Mel warping function function is defined as:

$$\tilde{\omega}_{Mel} = 2\pi \cdot 2595 \cdot \lg\left(1 + \frac{\omega}{2\pi \cdot 700\text{Hz}}\right). \tag{4.5}$$

The Mel warped magnitude spectrum is compressed by a filter bank to ≈ 20 components depending on the sample rate. The filter bank consists of equidistant triangular filters. The transfer function of a filter centered at $\tilde{f} = \tilde{\Omega}$ warped frequency is defined as follows:

$$H_{\tilde{\Omega}}(\tilde{\omega}) = \begin{cases} 0 & : \quad \tilde{\omega} - \tilde{\Omega} \leq -\frac{\tilde{B}}{2} \\ \frac{2}{\tilde{B}}(\tilde{\omega} - \tilde{\Omega}) + 1 & : \quad -\frac{\tilde{B}}{2} < \tilde{\omega} - \tilde{\Omega} \leq 0 \\ -\frac{2}{\tilde{B}}(\tilde{\omega} - \tilde{\Omega}) + 1 & : \quad 0 < \tilde{\omega} - \tilde{\Omega} \leq \frac{\tilde{B}}{2} \\ 0 & : \quad \tilde{\omega} - \tilde{\Omega} > \frac{\tilde{B}}{2} \end{cases} \tag{4.6}$$

where B defines the bandwidth of a filter. In this thesis, $\tilde{B} = 2\pi \cdot 268.258$ Mel-Hz has been used resulting in $N_{FB} = 15$ filters for 8kHz sample rate and in $N_{FB} = 20$ filters for 16kHz. The first filter is placed at a center frequency of $\frac{\tilde{B}}{2}$. The filters are placed at equidistant center frequencies. Distance between two center points is $\frac{\tilde{B}}{2}$. Hence, each filter overlaps with the predecessor and the successor filters. Although Mel warping and linear filter bank are two independent steps, a merged approach has significant advantages. Beside saving computation power, merging the warping into the filter bank algorithm avoids interpolation of the magnitude spectrum required when the Mel warping is implemented

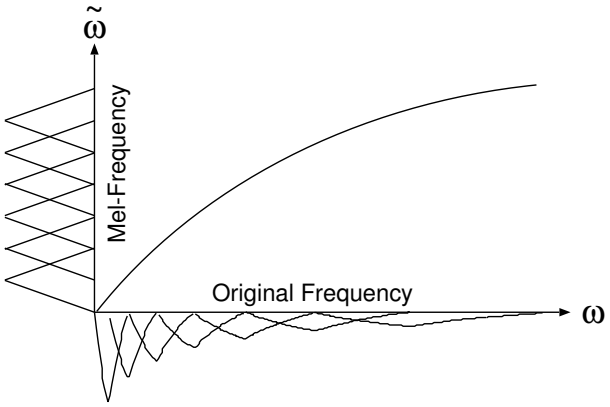


Figure 4.2: Linear filter bank in Mel-frequency domain and Mel warped filter bank in original frequency domain.

separately. The integration of the Mel warping is based on the idea of inverse warping the filters instead of the warping magnitude spectrum. Fig. 4.2 shows the example of an inverse warped filter bank consisting of seven filters. A more detailed discussion of the implementation of the integrated warping approach is presented in Section 4.4.1. For each filter, the output is the sum of the weighted spectral magnitudes. Logarithm is next applied followed by the Discrete Cosine Transform which decorrelates the filter bank outputs. In the RWTH implementation, a Discrete Cosine Transform is chosen which assumes that the filter bank output $F[n]$ is even about the point $N_{FB} - 0.5$ where N_{FB} denotes the number of filters:

$$c[k] = \sum_{n=0}^{N_{FB}-1} F[n] \cos\left(\pi \frac{n+0.5}{N_{FB}} k\right). \tag{4.7}$$

The optimal number of cepstrum coefficients depends on the corpus, see Table A.1. Subsequently, a normalization is carried out in order to account for different audio channels. Two types of normalization are applied: sentence-wise and session-wise. For sentence-wise recorded corpora, the cepstral mean and variance normalization is performed on whole sentences. In addition, the zeroth coefficient is shifted so that the maximum value within every sentence is zero (energy normalization). Session-wise recorded corpora consist of recordings containing several sequentially spoken sentences. For these corpora, cepstral mean normalization is carried out with a symmetric sliding window of 2s without variance and energy normalizations. In this manner, every 10ms a feature vector consisting of 16 normalized cepstrum coefficients is computed.

Baseline recognition tests using state-the-of-art acoustic features have been carried out on different corpora. Table 4.1 summarizes the results obtained by using among others the MFCC feature.

4.2 Perceptual Linear Predictive Analysis

Although the motivation of the Perceptual Linear Predictive (PLP) feature, proposed in [Hermansky 1990], is similar to the one of the MFCCs, there are major differences in extraction steps and in the recognition performance. As depicted on Fig. 4.3, every 10ms,

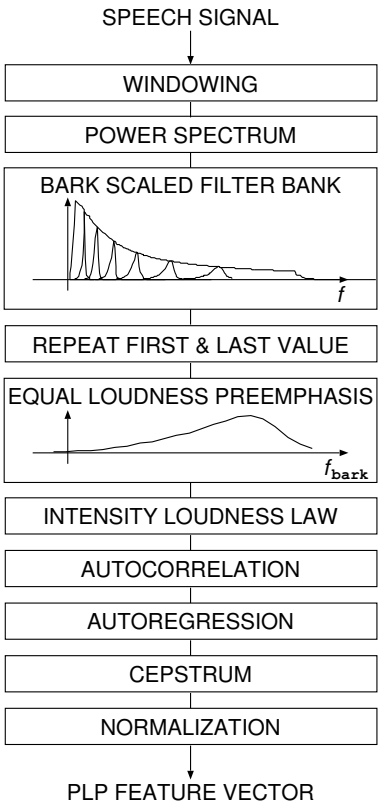


Figure 4.3: Block diagram of Perceptual Linear Predictive Analysis.

a Hamming window is directly applied to the speech signal omitting the preemphasis. Unlike in the MFCC method, a window length of 20ms is used. Comparative experiments have shown that on small- and on large-vocabulary tasks, 20ms long time frames yield significantly better results than using 25ms length. The short-term power spectrum is calculated by applying the FFT along with an appropriate zero padding. The necessary amount of zeros is determined in the same way as described in Section 4.1. In the next steps, first warping of the power spectrum is carried out followed by the application of a filter bank. In the PLP analysis, instead of the Mel warping function, the Bark warping is used:

$$g_{Bark}(\omega) = 2\pi \cdot 6 \cdot \ln\left(\frac{\omega}{2\pi \cdot 600\text{Hz}} + \sqrt{\left[\frac{\omega}{2\pi \cdot 600\text{Hz}}\right]^2 + 1}\right) \tag{4.8}$$

$$= 6\,\text{arcsinh}\left(\frac{\omega}{2\pi \cdot 600\text{Hz}}\right). \tag{4.9}$$

A filter bank consisting of trapezoidal filters is applied to the warped power spectrum. A trapezoidal filter with center frequency $\tilde{\omega} = \tilde{\Omega}$ is defined as follows:

$$H_{\tilde{\Omega}}(\tilde{\omega}) = \begin{cases} 0 & : \ \tilde{\omega} - \tilde{\Omega} \leq -2.5 \text{ Bark-Hz} \\ 10^{(\tilde{\omega}-\tilde{\Omega})+0.5} & : \ -2.5 \text{ Bark-Hz} < \tilde{\omega} - \tilde{\Omega} \leq -0.5 \text{ Bark-Hz} \\ 1 & : \ -0.5 \text{ Bark-Hz} < \tilde{\omega} - \tilde{\Omega} \leq 0.5 \text{ Bark-Hz} \\ 10^{-2.5[(\tilde{\omega}-\tilde{\Omega})-0.5]} & : \ 0.5 \text{ Bark-Hz} < \tilde{\omega} - \tilde{\Omega} \leq 1.3 \text{ Bark-Hz} \\ 0 & : \ \tilde{\omega} - \tilde{\Omega} > 1.3 \text{ Bark-Hz} \end{cases} \tag{4.10}$$

Filters are located at equidistant positions $\tilde{\Omega} = \tilde{n} \cdot \Delta\tilde{\omega}$ where $\tilde{n} = 1, ..., N_{FB}$. Number of filters N_{FB} is identical to the one used in the MFCC feature. For 8kHz sample, choosing $\Delta\tilde{\omega} = 0.973442$ results in 15 filter bank outputs, and for 16kHz sample rate, $\tilde{\Delta}\tilde{\omega} = 0.93853$ has to be chosen to obtain 20 filters. Bark warping is integrated into the filter bank in case of the PLP feature as well. Fig. 4.4 shows the example of an inverse warped filter bank consisting of seven filters. The filter bank is virtually extended by two more filters,

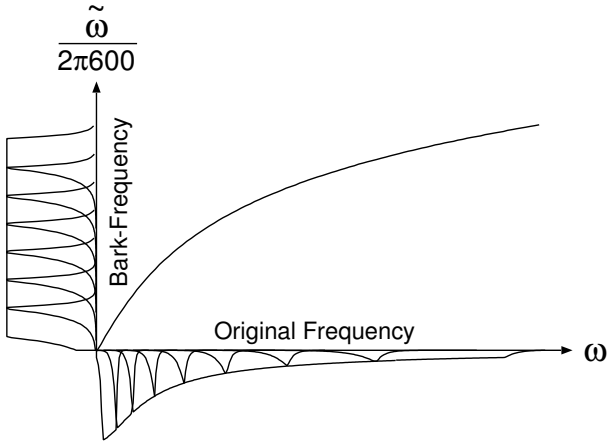


Figure 4.4: Trapezoidal filter bank in Bark-frequency domain and Bark-warped filter bank in original frequency domain. For sake of comparability, the warped frequency $\tilde{\omega}$ has been normalized by $2\pi \cdot 600$Bark-Hz similar to Eq. (4.8)

centered at frequency 0 and at sample-rate/2. Since these filters reach far beyond the valid frequency range, their output is discarded and replaced by the value of the right or left neighbor, respectively. Note, that through this, the number of filter bank outputs N_{FB} has been increased by two. Equal loudness preemphasis is applied to the extended filter bank outputs $F(\tilde{\Omega})$. The equal loudness preemphasis curve is given in form of the transfer function of a linear filter $l(\omega)$:

$$l(\omega) = \frac{\omega^4(\omega^2 + 5.68 \cdot 10^7)}{(\omega^2 + 6.3 \cdot 10^6)^2(\omega^2 + 3.8 \cdot 10^8)(\frac{\omega^6}{9.58 \cdot 10^26} + 1)} \tag{4.11}$$

Note that $\tilde{\Omega}$ has to be warped back to the original frequency domain ω since the equal loudness preemphasis curve is defined in that domain. Hence, the output of equal loudness

preemphasis $L(\tilde{\Omega})$ can be determined as follows:

$$L(\tilde{\Omega}) = F(\tilde{\Omega}) \cdot l(g_{Bark}^{-1}(\tilde{\Omega})). \tag{4.12}$$

Furthermore, for speech signals sampled at 8kHz, a simplified equal loudness preemphasis curve is used which ignores frequencies greater than 4kHz:

$$l_{4\text{kHz}}(\omega) = \frac{\omega^4(\omega^2 + 5.68 \cdot 10^7)}{(\omega^2 + 6.3 \cdot 10^6)^2(\omega^2 + 3.8 \cdot 10^8)}. \tag{4.13}$$

The equal loudness preemphasis is followed by the application of the intensity loudness law $I(\tilde{\Omega})$:

$$I(\tilde{\Omega}) = \sqrt[3]{L(\tilde{\Omega})}. \tag{4.14}$$

In the next stage of the algorithm, the cepstrum coefficients are derived from the all-poles approximation $\hat{I}(\tilde{\Omega})$ of the output of the intensity loudness law $I(\tilde{\Omega})$. First, autocorrelation coefficients $R[k]$ are calculated by applying the Inverse Discrete Fourier Transform to the output of the intensity loudness law $I(\tilde{n}\Delta\tilde{\omega}) = I[\tilde{n}]$. Since $I[\tilde{n}]$ is one half a symmetric function, the Inverse Discrete Fourier Transform can be calculated by applying a Discrete Cosine Transform which assumes that its input is even about the point $N_{FB} - 1$:

$$R[k] = \frac{1}{2}I[0] + (-1)^k \frac{1}{2}I[N_{FB} - 1] + \sum_{\tilde{n}=1}^{N_{FB}-2} I[\tilde{n}] \cos\left(\pi \frac{\tilde{n}}{N_{FB} - 1} k\right). \tag{4.15}$$

To obtain C cepstrum coefficients, the $C + 1$ autocorrelation coefficient are transformed to the gain $a[0]$ and to C autoregressive coefficients $a[\cdot]$ by using the Levinson-Durbin recursion. In this thesis, the Levinson-Durbin recursion published in [Rabiner and Schafer 1978] has been used with slight changes made by Hermansky and Morgan:

$E[0] = R[0];$
$\alpha[1][1] = k[1] = -\frac{R[1]}{R[0]};$
$E[1] = R[0] + R[1]k[1];$
FOR $i = 2, \ldots, C + 1$
    $k_[i] = -\frac{R[i] + \sum_{j=1}^{i-1} \alpha[j][i-1]R[i-j]}{E_[i-1]};$
    $\alpha[i][i] = k_[i];$
    FOR $j = 1, \ldots, (i - 1)$
        $\alpha[j][i] = \alpha[j][i - 1] + k[i]\alpha[i - j][i - 1];$
    $E_[i] = (1 - k[i]^2 E[i - 1];$

$a[0] = \sqrt{E[C + 1]};$
$a[\cdot] = \alpha[1, \ldots, C + 1][C + 1];$

Using the autoregressive coefficients $a[0,\ldots,C]$, the output of the intensity loudness law $I[\tilde{n}]$ is approximated by an all-poles model outputting $\hat{I}[\tilde{n}]$:

$$\hat{I}[\tilde{n}] = \sum_{i=1}^{C} a[i]\hat{I}[\tilde{n}-i] \tag{4.16}$$

$$I[0] = a[0]. \tag{4.17}$$

Instead of regenerating the smoothed all-poles approximation of the output of the intensity loudness law, the C cepstrum coefficients can be directly computed by applying a simple recursion:

$c[0] = 2\log(a[0]);$
$c[1] = -a[1];$
FOR $i = 2,\ldots,C-1$
$\quad c[i] = -\frac{1}{i}\left\{i\cdot a[i] + \sum_{k=1}^{i-1}(i-k)c[i-k]a[k]\right\};$

Note that the zeroth cepstrum coefficient is explicitly set to the logarithm of the square of the gain. Finally, a normalization is applied to the cepstrum coefficients as described in Section 4.1.

Recognition tests using the PLP feature have been carried out on two large-vocabulary corpora. As shown in Table 4.1, the PLP feature performs consistently worse than the other auditory motivated features, the MFCCs and MF-PLPs.

4.3 PLP Derived from Mel Scale Filter Bank

In this method, the MFCC and PLP techniques are merged into one algorithm generating the MF-PLP feature. When comparing the MFCC and PLP features, important differences can already be found in the first steps. Unlike the MFCCs, the PLP feature does not make use of the preemphasis of the speech signal. The MFCCs operate on the magnitude spectrum whereas the PLP extraction processes the power spectrum. At first view, the two methods use different warping function. However, plotting the functions in the common coordinate system shows that the differences are actually rather small, see Fig. 4.5. For sake of comparability, the Bark warping function has been scaled by an a such that $g_{Mel}(2\pi\cdot 8\text{kHz}) = a\cdot g_{Bark}(2\pi\cdot 8\text{kHz})$. Furthermore, the two feature extraction techniques differ by the choice of the filter function in the filter bank. The MFCC feature uses the rather simple triangular filter whereas the PLP feature is based on a trapezoidal one. As shown on Fig. 4.6, the first steps of the MF-PLP method until generating the output of the Mel scale triangular filter bank are mainly taken over from the MFCC algorithm making use of the preemphasis and the simple triangular filter bank. The only difference here compared to the MFCCs is that the filter bank is applied to the power spectrum instead of the magnitude spectrum.

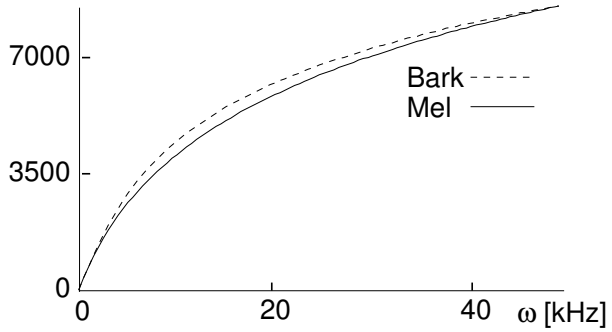


Figure 4.5: Mel $g_{Mel}(\omega)$ Eq. (4.5) and Bark $g_{Bark}(\omega)$ Eq. (4.8) warping functions. For sake of comparability, the Bark warping function has been scaled by an a such that $g_{Mel}(2\pi \cdot 8\text{kHz}) = a \cdot g_{Bark}(2\pi \cdot 8\text{kHz})$.

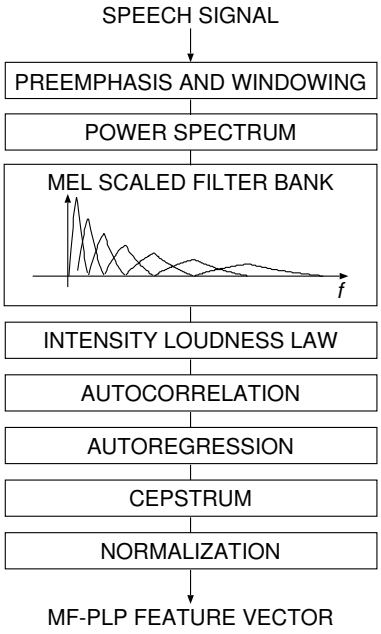


Figure 4.6: Block diagram of MF-PLP feature.

The second half of the MF-PLP signal analysis is motivated by the PLP algorithm. In the MF-PLP algorithm, two steps are abandoned from the PLP analysis: copying the outermost filters and the equal loudness preemphasis. The dynamic range of the filter bank outputs is compressed by the intensity loudness law. Fig 4.7 compares the intensity loudness law used in the PLP as well as in the MF-PLP analyzes with the logarithm used when extracting the MFCC feature. Although both functions share the same form, the logarithm compresses higher amplitude values stronger.
The cepstrum coefficients are calculated from the output of the intensity loudness law identically to the PLP algorithm as described in Section 4.2. The all-poles smoothing of the outputs of the intensity loudness law is carried out by determining the autocorrelation coefficients and applying the Levinson-Durbin recursion. The resulting autoregressive coefficients are transformed directly into cepstrum coefficients by using a simple recursion.

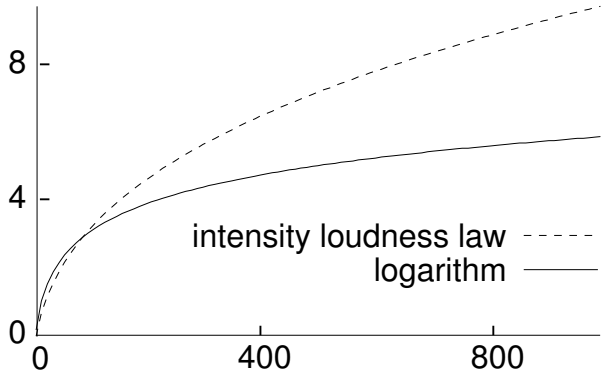


Figure 4.7: Comparison of dynamic range compression functions. The PLP feature uses the intensity loudness law $\sqrt[3]{x}$ whereas the MFCC feature uses the logarithm $\ln(x)$.

In comparison to the MFCC method, the major difference is that the MFCC method does not apply any smoothing to the output of the dynamic range compression.

Finally, a normalization is applied as described in Section 4.1. Recognition tests using the MF-PLP feature have been performed on two large-vocabulary corpora. As shown in Table 4.1, the MF-PLP feature performs consistently similar to the MFCCs.

4.4 Vocal Tract Length Normalization

A considerable part of the variability in the speech signal is caused by the speaker dependent vocal tract length. Vocal Tract Length Normalization tries to account for the effect of speaker specific vocal tract lengths by warping the frequency axis of the power spectrum of a speech segment.

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.

In this work, vocal tract length normalized features are used in combination with other auditory and articulatory motivated features. It is shown that even the performance of speaker adapted features can be improved by combining them with articulatory motivated features. Although vocal tract length normalization can be applied to any of the MFCC, PLP, and MF-PLP features, in this work, VTLN denotes the normalization of the MFCC feature. Baseline recognition results using solely the VTLN feature are summarized in Table 4.1.

The majority of the RWTH Vocal Tract Length Normalization system has been implemented by S. Molau. In his thesis [Molau 2003] a comprehensive description is provided over the implementation of frequency warping, training of warping factors, training of the normalized acoustic model, and recognition process using speaker normalization. In this work, improvements to Molau's VTLN system are presented

which aimed at simplifying the implementation and speeding up the training process. Professor S. Umesh has assisted this work by his new ideas and numerous experiments. The remaining part of this section is organized as follows: Implementation of the frequency warping is reviewed in Section 4.4.1. The optimization of the training procedure of warping factors is presented in Section 4.4.2 followed by the description of the training of the normalized acoustic model in Section 4.4.3. This section is closed by the summary of changes undertaken to optimize the integration of VTLN into the recognition process in Section 4.4.4.

Table 4.1: Word error rates for baseline acoustic features.

corpus	acoustic feature	error rates [%] del	ins	WER
SieTill	MFCC	0.3	0.5	1.8
VM II	MFCC	4.5	2.9	21.0
	VTLN	3.8	2.9	19.1
	MF-PLP	5.2	2.3	21.0
	PLP	5.9	2.3	21.4
EPPS	MFCC	4.3/3.8	1.4/1.7	14.7/15.3
	VTLN	4.3/3.7	1.3/1.5	14.2/14.1
	MF-PLP	4.2/3.7	1.5/1.7	14.8/15.3
	PLP	4.3/3.5	1.6/1.8	15.4/15.8

4.4.1 Implementation of Frequency Warping

In a Vocal Tract Length Normalization (VTLN) system, the frequency axis is warped twice during the calculation of the warped MFCC coefficients. First, a speaker dependent warping is carried out to account for speaker dependent vocal tract lengths. Different warping functions have been proposed to model speaker dependent changes of the frequency axis. A piecewise linear function was first tested in [Wegmann et al. 1996]. In [McDonough 1998], an all-pass transform has been proposed and in [Acero 1990, p.119], a bilinear transformation has been used to carry out the speaker dependent warping of the frequency axis. In the RWTH system, the piecewise linear function (Fig. 4.8) has been found to be the best choice with respect to word error rate. The piecewise linear function is defined for a normalized frequency ω in the following way:

$$\tilde{\omega}_\alpha = \begin{cases} \alpha\omega & : \quad \omega \leq \omega_0 \\ \alpha\omega_0 + \dfrac{\pi - \alpha\omega_0}{\pi - \omega_0}(\omega - \omega_0) & : \quad \omega > \omega_0 \end{cases} \tag{4.18}$$

where the position of ω_0 has been optimized empirically and it is chosen to be at $\omega_0 = \frac{7}{8} \cdot \pi$. After the VTLN warping, the standard Mel warping is applied carrying out the second warping step:

$$\tilde{\omega}_{Mel} = 2\pi \cdot 2595 \cdot \lg\left(1 + \frac{\omega}{2\pi \cdot 700\text{Hz}}\right). \tag{4.19}$$

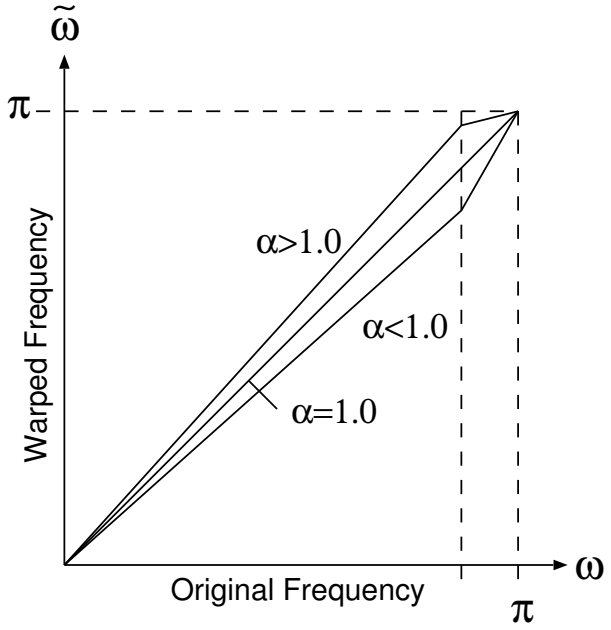


Figure 4.8: Piece-wise linear warping function with warping factor α.

Different implementations have been proposed for warping the frequency axis within the RWTH speech recognition group. The implementation in [Molau 2003] is based on separate nodes carrying out the two consecutive warping steps independently. In this setup, the filter bank consisting of equidistant triangular filters is applied to the warped magnitude spectrum. Fig. 4.9 illustrates this solution. This implementation can be improved at two stages:

- The two warping function can be nested yielding a single one. Nesting two functions does not change the properties of the underlying functions with respect to the inverse and the derivative. If the two functions were invertible, the result of nesting them remains invertible. Derivation of nested functions can be automatized by implementing the chain rule. Namely, if the derivative functions of the two underlying functions are known, the derivative of the resulting nested function can be obtained without any manual effort. Hence, the implementation of warping does not depend on the function. Even nested functions can be seen as a single function thus they can be applied without changing the warping algorithm.

- Direct implementation of warping requires interpolation over either the original or the warped frequency axis. The interpolation is inevitable because the warped axis is sampled at equidistant samples. For nonlinear warping function, it cannot be guarantied that a sample of the original frequency axis falls automatically on a sample of the warped frequency axis. One of the most attractive solutions is to integrate the warping into the filter bank. This solutions avoids the separate warping step making the interpolation unnecessary and saving computational power at the same time.

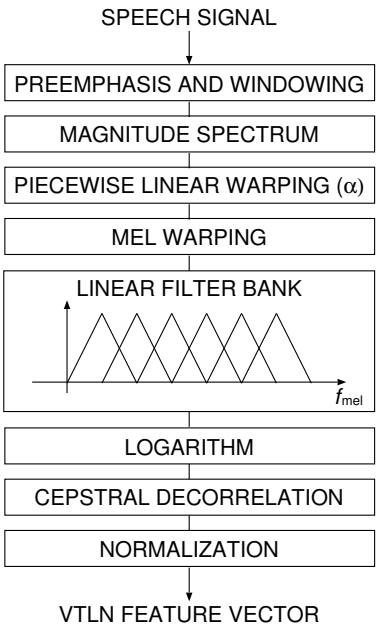


Figure 4.9: Block diagram of vocal tract length normalized MFCC feature (VTLN) using separate warping nodes.

Both improvements have been addressed in [Pitz and Ney 2003], however in a VTLN system which has discarded the filter bank from the MFCC data flow. Nested warping functions have been directly integrated into the discrete cosine transform to perform the Mel and VTLN warping in one step. In [Umesh et al. 2005], the necessity of the application of a filter bank is pointed out. Consequently, the warping is integrated into the filter bank rather than into the discrete cosine transform. In the following, the implementation is reviewed which has been used in this paper as well as in feature combination experiments presented in this thesis. Note that the derivation is carried out in the continuous domain for sake of simplicity. Discretization of results is performed at the end to provide the equations which have been used throughout the VTLN experiments.

A warping of the frequency axis ω is given by the function $g(\omega)$. The resulting warped axis is denoted by $\tilde{\omega}$:

$$\tilde{\omega} \quad = \quad g(\omega). \tag{4.20}$$

Correspondingly, the warping of the magnitude spectrum $X(\omega)$ is carried out so that the following equalities holds:

$$\tilde{X}(\tilde{\omega}) \quad = \quad \tilde{X}(g(\omega)) = X(\omega). \tag{4.21}$$

With other words, the value of the magnitude spectrum $X(\omega)$ at ω is copied to the warped frequency $\tilde{\omega} = g(\omega)$ making up the warped magnitude spectrum $\tilde{X}(\tilde{\omega})$ at the warped frequency $\tilde{\omega}$.

In the standard MFCC implementation, the filter bank is applied to the warped magnitude spectrum as shown on Fig. 4.9. The filter bank applied to the MFCC feature

consists of equidistant linear triangular filters $\tilde{H}(\tilde{\omega})$:

$$\tilde{H}(\tilde{\omega}) = \begin{cases} 0 & : \quad \tilde{\omega} \leq -\frac{\tilde{B}}{2} \\ \frac{2}{\tilde{B}}\tilde{\omega} + 1 & : \quad -\frac{\tilde{B}}{2} < \tilde{\omega} \leq 0 \\ -\frac{2}{\tilde{B}}\tilde{\omega} + 1 & : \quad 0 < \tilde{\omega} \leq \frac{\tilde{B}}{2} \\ 0 & : \quad \tilde{\omega} > \frac{\tilde{B}}{2} \end{cases} \tag{4.22}$$

where B denotes the bandwidth of a filter. Output of the filter bank $F(\tilde{\Omega})$ is obtained by calculating the convolution of warped magnitude spectrum $\tilde{X}(\tilde{\omega})$ and the transfer function of a filter $\tilde{H}(\tilde{\omega})$:

$$F(\tilde{\Omega}) = \int_{\tilde{\omega}=-\infty}^{\infty} \tilde{X}(\tilde{\omega})\tilde{H}(\tilde{\omega} - \tilde{\Omega})\, d\tilde{\omega}. \tag{4.23}$$

$F(\tilde{\Omega})$ is evaluated at ≈ 20 equidistant samples. In case of triangular filters, the filters are placed at distances of $\frac{B}{2}$.

The goal of the integrated warping approach is to calculate the filter bank outputs directly on the unwarped magnitude spectrum $X(\omega)$ rather than on the warped one. In this way, the interpolation required when calculating the warped magnitude spectrum can be eliminated. Hence the goal of the following derivation is, to calculate the filter bank output $F(\tilde{\Omega})$ directly on $X(\omega)$ and at the same to carry out the warping within the convolution. In the first step, Eq. (4.20) is applied to the definition of the filter bank output Eq. (4.23):

$$F(\tilde{\Omega}) = \int_{\omega=-\infty}^{\infty} \tilde{X}(g(\omega))\tilde{H}(g(\omega) - \tilde{\Omega})\frac{\partial g(\omega)}{\partial \omega} d\omega \tag{4.24}$$

Note that derivative of the warping function is required in order to change the integration variable from $\tilde{\omega}$ to ω. Applying the definition of warping Eq. (4.21) yields the first form where $F(\tilde{\Omega})$ is already calculated on the original unwarped magnitude spectrum $X(\omega)$:

$$F(\tilde{\Omega}) \overset{\tilde{X}(g(\omega))=X(\omega)}{=} \int_{\omega=-\infty}^{\infty} X(\omega)\tilde{H}(g(\omega) - \tilde{\Omega})\frac{\partial g(\omega)}{\partial \omega} d\omega. \tag{4.25}$$

In the following, the notion of a *warped filter* is introduced to define the basis of an efficient implementation of the integrated warping approach. The idea is based on the duality between the magnitude spectrum and the filter bank. In the original approach, the unwarped filters have been applied to the warped magnitude spectrum. In Eq. (4.25), the filter bank output is already calculated on the original unwarped magnitude spectrum. Hence, the duality indicates that the pair of the unwarped magnitude spectrum is a warped filter. Although a more correct name would be the *inverse-warped filter*, the terminology *warped filter* is used in most of the cases. The warped filter H is defined as follows:

$$H_{\tilde{\Omega}}(\omega) = H(g^{-1}(\tilde{\omega} - \tilde{\Omega})) = \tilde{H}(\tilde{\omega} - \tilde{\Omega}). \tag{4.26}$$

Note that inverse warping of the filter $\tilde{H}(\tilde{\omega})$ cannot be carried out independently of $\tilde{\Omega}$ since any nonlinear warping function is allowed. Nevertheless, for a given $\tilde{\Omega}$ the warped filter can be precalculated for equidistant samples $H_{\tilde{\Omega}}(n\Delta\omega)$. Seven examples of $H_{\tilde{\Omega}}(\omega)$ are shown on Fig. 4.10 for the piecewise linear warping ($\alpha = 0.9$) followed by the Mel warping. After defining the notion of the warped filter, the filter bank implementation can almost be written in the form of a convolution over the original frequency axis. However, the convolution parameter $\tilde{\Omega}$ changes the function H in a nonlinear manner:

$$F(\tilde{\Omega}) = \int_{\omega=-\infty}^{\infty} X(\omega) \left[H_{\tilde{\Omega}}(\omega) \frac{\partial g(\omega)}{\partial \omega} \right] d\omega. \tag{4.27}$$

In order to implement the integrated approach, a discretization of Eq. (4.27) is carried out in the following. The filter bank is evaluated at equidistant samples $\tilde{n}\Delta\tilde{\omega}$. The magnitude spectrum is given at discrete samples $X[n]$ where $n = 0, ..., \frac{N}{2}$ and N is the number of FFT points. The warped filter along with the derivative of the waring function is queried at equidistant values $n\Delta\omega$ where $\Delta\omega = \frac{f_s}{N}$ and f_s denotes the sample rate. The warped filters are precalculated for the different values of $\tilde{n}$.

$$F[\tilde{n}] = \sum_{n=0}^{\frac{N}{2}} X[n] \left[H_{\tilde{n}\Delta\tilde{\omega}}(n\Delta\omega) \frac{\partial g(n\Delta\omega)}{\partial \omega} \right] \Delta\omega \tag{4.28}$$

In the RWTH implementation two simplifications have been made yielding the final form of the warped filter bank Eq. (4.29): On the one hand, the multiplication with the differential unit $\Delta\omega$ has been neglected. On the other hand, the distance between two samples of the warped spectrum $\Delta\tilde{\omega}$ is chosen to be equal to the one of the original magnitude spectrum $\Delta\omega$.

$$F[\tilde{n}] = \sum_{n=0}^{\frac{N}{2}} X[n] \left[H_{\tilde{n}\Delta\omega}(n\Delta\omega) \frac{\partial g(n\Delta\omega)}{\partial \omega} \right] \tag{4.29}$$

In summary, the integrated warping approach makes use of two facts: On the one hand, warping function $g(\omega)$ is given in a continuous form thus it can be queried at any value of ω. On the other hand, the original magnitude spectrum $X(\omega)$ is given at discrete samples $X[n]$ thus the sum can only make use of these values. Hence, the interpolation of magnitude spectrum becomes unnecessary in this approach because the filter bank is warped which is given in the form of continuous functions rather the discrete magnitude spectrum. The block diagram of the vocal tract length normalized MFCC feature using integrated warping is shown on Fig. 4.10.

In the following, important remarks on the integrated warping approach are summarized with regard to its implementation and application:

- The filter bank given in Eq. (4.29) operates on the discrete spectrum. If the second term of the summand is precalculated for some $\tilde{n}$ and n pairs, the formula reduces to

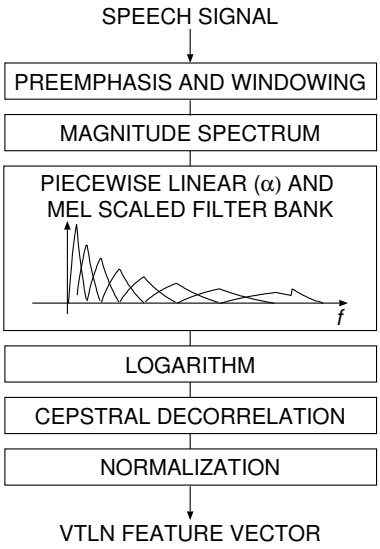


Figure 4.10: Block diagram of vocal tract length normalized MFCC feature (VTLN) using integrated warping.

a matrix-vector multiplication. The matrix $A_{\tilde{n}n}$ containing the precalculated values is defined as:

$$A_{\tilde{n}n} = \left[H_{\tilde{n}\Delta\omega}(n\Delta\omega) \frac{\partial g(n\Delta\omega)}{\partial\omega} \right] \tag{4.30}$$

and the vector contains the unwarped magnitude spectrum. Hence, the application of a filter bank with integrated warping reduces to two kinds of operations. First, before the signal analysis starts, the matrix $A_{\tilde{n}n}$ is precalculated. Then, for every time frame a simple matrix-vector multiplication is carried out. Note that the matrix contains mainly zero elements thus the time spent on multiplications with zeros can be saved by using an appropriate data structure.

- The derivative of the warping function adjusts the infinitesimal quantity $d\omega$ when changing the integration variable from $\tilde{\omega}$ to ω. If the filter bandwidth is not large and the warping function is not changing fast, the derivative can be approximated by a constant depending on the position of the filter $\tilde{\Omega}$. Hence, the derivative can be approximated by global diagonal matrix scaling the output of the filter bank. Consequently, under the aforementioned constraints, the derivative does not yield any classification information. Comparative experiments have verified this statement. Nevertheless, the RWTH implementation includes the derivative of the warping function since it is only queried in the precalculation phases i.e. it does not effect the effectiveness of the transformation of time frames.

- The algorithm presented in this section is general so much so that the signal analysis block diagram presented on Fig. 4.10 can calculate the MFCC and VTLN features without any change in the implementation. The only difference between the two systems is how the warping function is defined.

4.4.2 Training of Warping Factors

In this work, maximum likelihood estimation has been applied to determine the warping factors. A grid search optimization is carried out for different discrete values of the warping factor α because a closed-form solution of the optimization criterion is not obvious [Lee and Rose 1996]. In [Lööf et al. 2006], a new method is published to speeding up the grid search. Due to the equivalence of frequency warping and linear transformation of cepstral coefficients presented in [Umesh et al. 2005], warping factors can be efficiently estimated by accumulating the sufficient statistics for linear transformation estimation and searching the constrained space of transformations given by the explicit mapping between warping factors and linear transformation matrices. Nevertheless, the standard warping function estimation has been used in this work.

In this section, changes made to the VTLN system described in [Molau 2003] are worked out without presenting the complete warping factor estimation procedure. In the RWTH VTLN system, a single densities based emission model is used in the grid search to calculate the acoustic probabilities. The use of single densities has been found to be beneficial during the grid search process since this simple emission model does not account for speaker dependent variations in the speech signal. In the implementation of Molau, the emission model has been trained on LDA transformed acoustic features. However, the LDA matrix has not been reestimated when evaluating different warping factors.

In this thesis, the single densities based emission model is trained on a simplified signal analysis. The LDA transformed features have been replaced by dynamic features. The MFCC coefficients have been augmented with their first derivative and the with the second derivative of the first MFCC coefficient. Note that the derivatives have been calculated over 5 successive time frames using a linear regression based method. Consequently, the acoustic model used in the warping factor estimation has been trained in the following way:

1. Generation of frame energy measures.
2. Generation of a start alignment by using linear segmentation based on frame energy measures.
3. Generation of MFCC coefficients augmented with the first derivative and the second derivative of the first MFCC coefficient.
4. Improvement of the start alignment by using a monophone single density acoustic model (10× alignment).
5. Estimation of triphone state tying tree (Classification and Regression Tree - CART).
6. Training of single densities based emission model (10× alignment).

Comparative recognition tests have not found any difference in WER between using Molau's LDA based or the simplified dynamic features based acoustic models. One of the major change is that that the LDA has been completely removed from the warping factor estimation simplifying the underlying acoustic model. A straightforward extension of the simplification process would be to discard even the dynamic features.

4.4.3 Training of Acoustic Model

One of the important results summarized in [Molau 2003] is that speaker normalization can be best applied in training as well as in recognition. Hence, the VTLN acoustic model is trained on normalized MFCC feature. Beside using a normalized acoustic feature, the training procedure is exactly the same as when applied to the MFCC feature. In the first step of the training, the normalized feature vectors are stored in a file cache. The subsequent training steps are using this feature cache independently of the concrete feature type stored in the cache.

The implementation of the training procedure has only been changed slightly compared to Molau's system. The initialization steps have been simplified in the system used in this thesis. Molau's acoustic training did not utilize the well trained single densities based acoustic model used in warping factor estimation, see in Section 4.4.2. The training has been started with a start alignment generated by an optimized MFCC based system. Molau's training procedure can be summarized as follows:

1. Generation of VTLN normalized acoustic feature vectors.
2. Estimation of the first CART tree using only the 16 MFCC coefficients and based on an optimized start alignment generated by using solely the MFCC feature.
3. Estimation of the first LDA matrix base on the first CART tree.
4. Estimation of the second CART tree using MFCC vectors transformed by the first LDA matrix.
5. Estimation of the second LDA matrix based on the second CART tree.
6. Training of the emission GMMs including 7 splits.

In this thesis, the initialization steps of the acoustic training have been simplified. Instead of estimating the first CART tree solely on the MFCC coefficients, the CART tree estimated during the single density training is reused. The training is constructed of the following steps:

1. Generation of VTLN normalized acoustic feature vectors.
2. Estimation of the first LDA matrix base on the CART tree and the start alignment produced by the single density training described in Section 4.4.2.
3. Estimation of a CART tree using MFCC vectors transformed by the first LDA matrix.
4. Estimation of the second LDA matrix based on the newly estimated CART tree.
5. Training of the emission GMMs like in Molau's system.

In summary, the training procedure of the VTLN normalized acoustic model has been taken over from Molau's system. Few changes have been made to the initialization part which reuse well trained parts of the single densities based acoustic model used in warping factor estimation.

4.4.4 Recognition Using VTLN

In [Molau 2003], two fundamentally different methods have been proposed for warping factor estimation during recognition. The first method reuses the warping factor estimation technique applied to the training corpus. In a two-pass recognition system, an orthography of the spoken utterance is generated first. The warping factor of the current speaker is estimated using a HMM structure which has been constructed on the basis of the recognized word sequence. In the second recognition pass, acoustic features are normalized by using the estimated warping factor. The second method described in [Welling et al. 2002] enables a single-pass recognition process. The warping factors are recognized based on scores which are delivered by GMMs trained text-independently to distinguish the predefined warping factors. The warping factors share an equal a-priori probability. After a simple recognition system has output a warping factor for the current speaker, the recognition process can be started by feeding in the normalized features.

A comparison of warping factor estimation methods has been carried out in [Molau 2003]. No significant differences have been found thus Molau recommends the GMM based one-pass VTLN system. The comparison has addressed a third method as well: supervised warping factor estimation. In this case, the two-pass system is sped up by discarding the first recognition pass and directly using the reference transcription of the test corpus as bases for the HMM structure. Results obtained by using the supervised method can be seen as lower bound for WERs obtained by using different one- or two-pass warping factor estimation methods. Also note that only slight or insignificant differences can been found in WER when comparing the supervised estimation to the one- and two-pass methods. In this thesis, supervised warping factor estimation has been applied to the *VM II* corpus for sake of faster recognition passes, and text-independent GMMs have been used on the *EPPS* corpus.

Similar to the simplification of the warping factor estimation in Section 4.4.2, the estimation of the text-independent GMMs have been reviewed as well in the scope of this thesis. In Molau's VTLN system, the text-independent GMMs have been trained on LDA transformed MFCC feature vectors rather than directly on the speaker independent MFCC feature. In [Molau 2003], no clear explanation can be found what advantage can be taken of the LDA transform. A further unclear detail is why a LDA matrix trained on the VTLN normalized MFCC feature has been used instead of a matrix trained directly on the MFCC feature. In this thesis, the one-pass VTLN recognition system has used a simplified feature extraction during warping factor estimation. The GMMs have been trained using the standard dynamic features i.e. MFCC feature coefficients have been augmented by their first derivative and by the second derivative of the first MFCC coefficient. The simplified feature extraction setup has not yield any degradation in WER on the *VM II* and *EPPS* corpora when compared to using LDA transformed feature vectors. Similar to the simplified warping factor estimation process, a straightforward continuation of this work would be to train the warping factor dependent GMMs directly on the MFCC coefficients without any dynamic feature.

4.5 Voicing Feature

One of the most intensively researched articulatory motivated features is the voicing feature. Voiced and unvoiced sounds form two complementary classes. Therefore, a method explicitly extracting the degree of voicing from a speech segment can be expected to improve discrimination of phonemes and consequently to improve recognition results. In rule-based speech recognition systems, voiced-unvoiced detection has been used as one of the acoustical features. In [Atal and Rabiner 1976], a voiced-unvoiced-silence detection algorithm is proposed using a multi-pass statistical approach. A voicing measure instead of a voiced-unvoiced decision is described in [Thomson and Chengalvarayan 1998]. The authors presented results obtained by using an autocorrelation based voicing measure along with liftered cepstral coefficients. Jitter of the extracted fundamental frequencies has been used as third acoustic feature. Using the concatenated features, a large relative improvement in WER has been achieved by applying discriminative training. A recently developed voicing feature has been published in [Graciarena et al. 2004]. The entropy of the high order cepstrum is used to extract voicing information. Recognition tests have shown significant improvement in WER when the entropy based voicing feature has been combined with an autocorrelation based voicing feature and with the MFCC feature.

In this Section, three different voicing extraction method are presented. Here, the goal was to produce a continuous measure representing the degree of vibration of the vocal cords instead of the implementation of a voiced-unvoiced decision algorithm. The oscillation of the vocal chords produces quasi periodic segments in the speech signal. Common motivation of the presented voicing extraction methods is to quantify this periodicity. In Section 4.5.1, a Harmonic Product Spectrum based method is presented which measures the periodicity of a time frame in the frequency domain. An autocorrelation based voicing measure is described in Section 4.5.2 and an Average Magnitude Difference based in Section 4.5.3. Both methods extract the voicing measure in the time domain.

The rest of this Section is organized as follows. The three methods are compared on a theoretical level followed by experimental results in Section 4.5.4. The autocorrelation based voicing measure is discussed in Section 4.5.5. Finally, recognition results are presented in Section 4.5.6 which have tested the autocorrelation based voicing measure on three different corpora and in combination with several baseline features.

4.5.1 Harmonic Product Spectrum

The magnitude spectrum of voiced sounds shows sharp peaks that occur at integer multiples of the fundamental frequency. This fact serves as basis for the method Harmonic Product Spectrum (HPS) [Rabiner and Schafer 1979]. In [Zolnay et al. 2002], a voicing measure is derived from the harmonic product spectrum. In the following, this method is presented. The harmonic product spectrum $P^{[t]}(n)$ for some time frame t and discrete frequency n is the product of R frequency-shrunken replicas of the magnitude spectrum

$|X^{[t]}(e^{j\frac{2\pi}{N}n})|$:

$$P^{[t]}(n) = \sqrt[R]{\prod_{r=1}^{R} |X^{[t]}(e^{j\frac{2\pi}{N}nr})|} \tag{4.31}$$

N is the number of FFT points and $R = \lfloor N/2n \rfloor$ is the maximum shrinkage of magnitude spectrum which can still provide a magnitude value on discrete frequency n. An example is shown on Fig. 4.11. The first diagram shows the magnitude spectrum of a quasi periodic frame from an utterance of the sound /a/. The next two diagrams picture the same magnitude spectrum shrunken by the factor of 2 and 3. The last diagram shows the harmonic product spectrum obtained by applying Eq. (4.31).

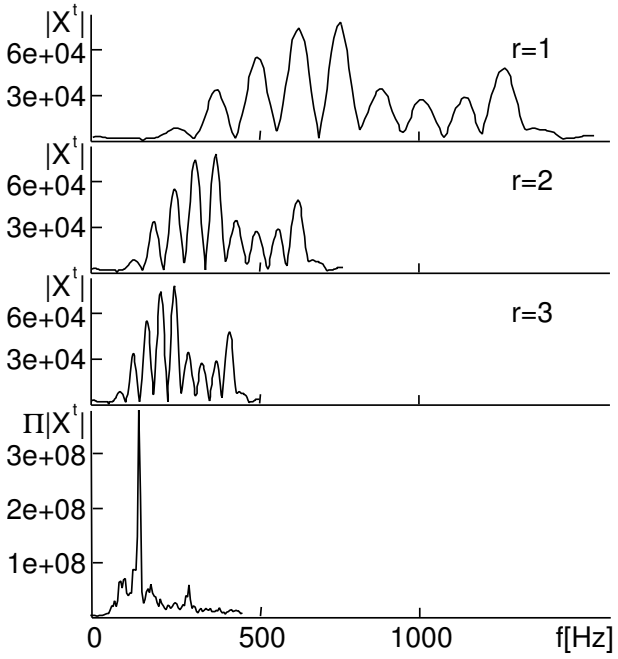


Figure 4.11: Shrunken magnitude spectra and the harmonic product spectrum of a 40ms long time frame from the middle of an utterance of the sound /a/.

The motivation for using the product spectrum is that for periodic signals, shrinking the frequency axis by integer factors forces the harmonics to coincide at the fundamental frequency and at its nearby harmonics. Since the magnitude spectrum of a periodic signal is zero between the harmonics, the product of shrunken magnitude spectra cancels out all the magnitude values which fall between two harmonics of the fundamental frequency. In ideal case, the harmonic product spectrum results in high peaks at the fundamental frequency and at its nearby harmonics and it is zero otherwise. Note that since speech analysis is based on short-time Fourier analysis and even voiced sounds are only quasi periodic, the harmonic product spectrum is small but not exactly zero between the harmonics of the fundamental frequency and its peaks are not always obvious. Fig. 4.12 shows the harmonic product spectra of two time frames extracted from the same utterance of the sound /a/. On the left hand side, an obvious peak can be found indicating that the time frame is voiced. The harmonic product spectrum depicted on the right hand side

has been calculated for a time frame at the end of the utterance. The voicing is decaying to the end resulting in a less obvious peak of the harmonic product spectrum.

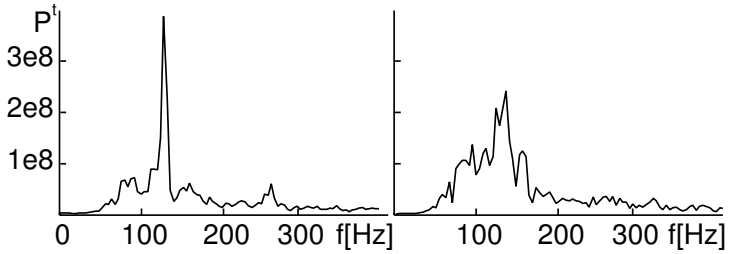


Figure 4.12: Harmonic product spectrum of two 40ms long time frames from the middle (left) and from the end (right) of an utterance of the sound /a/.

The goal of voicing extraction is to produce a bounded value describing how voiced the current time frame is. In the following, a measure will be presented which captures the peak structure of the harmonic product spectrum. Voiced time frames exhibit a sharp maxima whereas unvoiced time frames have no clear peak structure and the maxima of the harmonic product spectrum is typically flat. The measure $\tilde{v}_{HPS}^{[t]}$ evaluates the maximum magnitude value of the harmonic product spectrum. It is defined as the ratio of the maximum value and the geometric mean of the neighboring values without the maximum:

$$n_{max}^{[t]} = \underset{80\text{Hz}\cdot\frac{N}{f_s}\leq n\leq 400\text{Hz}\cdot\frac{N}{f_s}}{\operatorname{argmax}} P^{[t]}(n), \tag{4.32}$$

$$\tilde{v}_{HPS}^{[t]} = \frac{P^{[t]}(n_{max}^{[t]})}{\sqrt[2W]{\prod_n P^{[t]}(n)}}, \tag{4.33}$$

where f_s is the sample rate and N is the number of FFT points. The maximum of the harmonic product spectrum n_{max} is searched only in the interval of practically relevant pitches [80Hz..400Hz]. The geometric mean is calculated over the neighborhood of n_{max}: n runs from $n_{max} - W$ to $n_{max} + W$ excluding n_{max}. The minimum pitch and thus the minimum distance between two harmonics is about 80Hz. Therefore, the size of the neighborhood is set to $W = 70\text{Hz} \cdot \frac{N}{f_s}$ to avoid peaks of the neighboring harmonics being included in the average. Typical values are $1 \leq \tilde{v}_{HPS}^{[t]} < 4$. Values $\tilde{v}_{HPS}^{[t]} > 2$ are cut to 2 since they obviously indicate a voiced segment:

$$v_{HPS}^{[t]} = \min\{2, \tilde{v}_{HPS}^{[t]}\} - 1. \tag{4.34}$$

Fig. 4.13 summarizes the necessary steps to extract the voicing measure. A Hamming window is applied every 10ms to speech segments of 40ms length. The window length has been empirically optimized on small- and large-vocabulary corpora. (More details on the optimization of the voicing extraction can be found in Section 4.5.5.) After windowing, the magnitude spectrum is determined. To increase the frequency resolution and thus to increase the number of magnitude values between two harmonics, a 2048-point FFT is computed with appropriate zero padding. The harmonic product spectrum

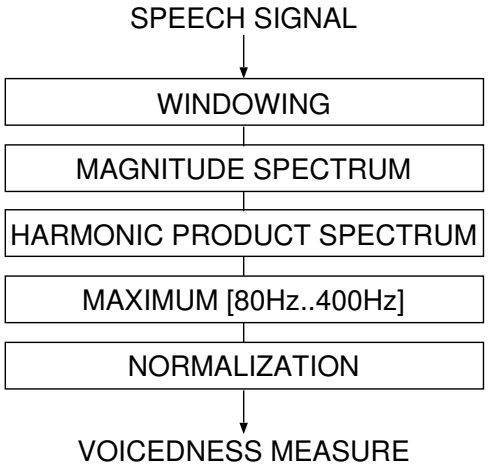


Figure 4.13: Block diagram of the harmonic product spectrum based voicing measure.

is calculated using the logarithm of the magnitude spectrum to compress the dynamic range of the speech signal. After calculating the harmonic product spectrum for the interval [80Hz .. 400Hz], a simple linear search is used to find the maximal value. The maximum is normalized by applying first Eq. (4.33) followed by Eq. (4.34). In this way, an one-dimensional voicing feature is generated every 10ms.

Finally, Fig. 4.14 depicts histograms of $v_{HPS}^{[t]}$ on voiced-unvoiced sound pairs. The plosive sound pair /g/-/k/ and the fricative sound pair /v/-/f/ are compared which phonetically differ only by the type of excitation (i.e. state of the vocal cords). The histogram of a given phoneme has been estimated on time frames aligned to any of the states of the triphones with the given phoneme as central one. Note that here the major difference between the histograms is the high peak at value 1 for voiced sounds.

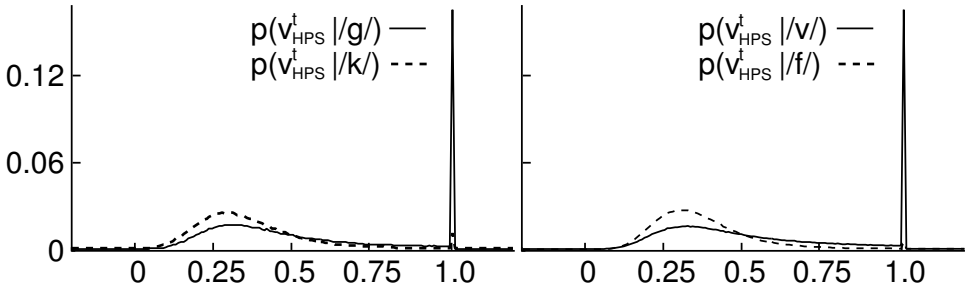


Figure 4.14: Histograms of the measure $v_{HPS}^{[t]}$ estimated on the *VM II* corpus. Left: voiced plosive sound /g/ and its unvoiced counterpart /k/. Right: voiced fricative sound /v/ and its unvoiced counterpart /f/.

4.5.2 Autocorrelation

In the following, the extraction of an autocorrelation (AC) based voicing measure is described. The voicing measure is calculated for short time frames instead of complete sounds. An example is shown on Fig. 4.15. The time frame can be seen as an infinite

long signal which has values other than zero only within the interval $[t..t+W]$ where W is the window length. Autocorrelation $R^{[t]}(\tau)$ expresses the similarity between the time

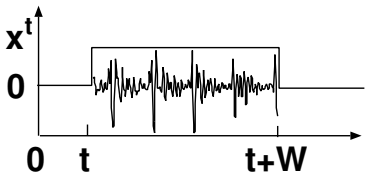


Figure 4.15: Windowed time frame of a voiced speech signal.

frame $x^{[t]}(\nu)$ and its copy shifted by τ. Assume the autocorrelation $R^{[t]}(\tau)$ for some time frame t and a shift τ:

$$R^{[t]}(\tau) = \frac{1}{W} \sum_{\nu=0}^{W-\tau-1} x^{[t]}(\nu)\ x^{[t]}(\nu+\tau). \tag{4.35}$$

The autocorrelation attains its maximum at $\tau = 0$. For periodic signals with a frequency f, the autocorrelation attains its maximum not only at $\tau = 0$ but also at $\tau = \frac{k}{f}$ $k = 0, \pm1, \pm2, ...$ integer multiples of the period. Hence, the autocorrelation of voiced time frames obtains values close to $R^{[t]}(0)$ at lags of integer multiples of the fundamental frequency. Therefore, a maxima in the range of practically relevant pitches of a value close to $R^{[t]}(0)$ is a strong indication of periodicity and consequently of voicing.

The straightforward application of Eq. (4.35) results in decreasing number of non-zero summands as τ approaches W. To account for the deceasing amount of summands, the unbiased estimate of the autocorrelation $\tilde{R}^{[t]}(\tau)$ has been used:

$$\tilde{R}^{[t]}(\tau) = \frac{1}{W-\tau} \sum_{\nu=0}^{W-\tau-1} x^{[t]}(\nu)\ x^{[t]}(\nu+\tau). \tag{4.36}$$

Note that the unbiased estimate does not necessarily share the properties of the autocorrelation defined in Eq. (4.35). Theoretically, it does not necessarily attains its maximum at lag $\tau = 0$. Nevertheless, unbiased estimates with a maximum at $\tau \neq 0$ appear negligibly seldom in practice. Fig. 4.16 shows the unbiased estimate of the autocorrelation of two time frames extracted from the same utterance of the sound /a/. On the left hand side, obvious equidistant peaks of values close to $\tilde{R}^{[t]}(0)$ indicate that the time frame is voiced. The autocorrelation depicted on the right hand side has been calculated for a time frame at the end of the utterance. The voicing is decaying to the end resulting in a less obvious and low peaks.

In order to produce a bounded measure of voicing, the autocorrelation is divided by $\tilde{R}^{[t]}(0)$. The resulting function provides values mainly in the interval $[-1..1]$ nevertheless because of the unbiased estimate, theoretically any value is possible. The voicing measure $v_{AC}^{[t]}$ is thus the maximum of the normalized autocorrelation in the interval of practically relevant pitch periods [2.5ms..12.5ms]:

$$v_{AC}^{[t]} = \frac{\max\limits_{2.5\text{ms}\cdot f_s \leq \tau \leq 12.5\text{ms}\cdot f_s} \tilde{R}^{[t]}(\tau)}{\tilde{R}^{[t]}(0)} \tag{4.37}$$

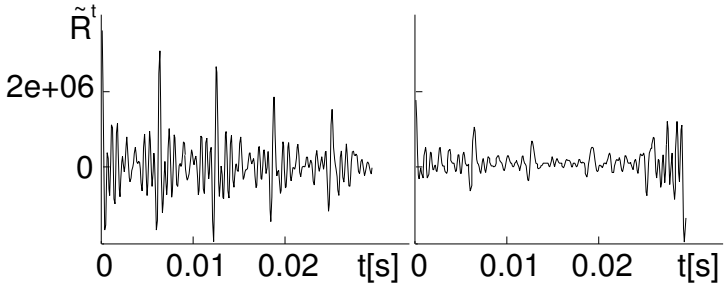


Figure 4.16: Unbiased estimate of the autocorrelation of two 40ms long time frames from the middle (left) and from the end (right) of an utterance of the sound /a/.

where f_s denotes the sample rate. Values of $v_{AC}^{[t]}$ close to 1 indicate voicing, values close to 0 indicate voiceless time frames. Fig. 4.17 summarizes the necessary steps to extract the voicing measure. The autocorrelation function is determined every 10ms on speech

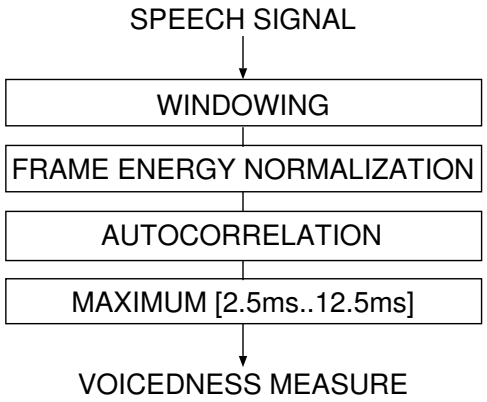


Figure 4.17: Block diagram of the autocorrelation based voicing measure.

segments of 40ms length. The window length has been empirically optimized on small- and large-vocabulary corpora. The optimal value of 40ms corresponds to results published formerly, see chapter Pitch Detection in [Rabiner and Schafer 1979]. The frame energy normalization ensures that $\tilde{R}^{[t]}(0) \equiv 1$ thus the division in Eq. (4.37) can be omitted. After calculating the unbiased autocorrelation for the discrete lags $[2.5\text{ms} \cdot f_s \,..\, 12.5\text{ms} \cdot f_s]$, a simple linear search is used to find the maximal value. In this way, an one-dimensional voicing feature is generated every 10ms.

Finally, Fig. 4.18 compares histograms of $v_{AC}^{[t]}$ on voiced-unvoiced sound pairs. The histograms have been estimated in the same way as described in Section 4.5.1. The voiced histograms are well separated form the unvoiced ones. Hence, the autocorrelation based voicing measure can lead to better discrimination of voiced-unvoiced sounds and consequently to better recognition results. Recognition tests presented in Section 4.5.6 confirm this observation on different corpora.

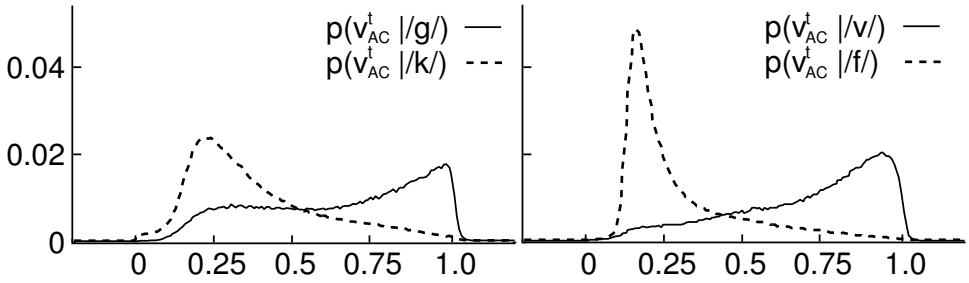


Figure 4.18: Histograms of the measure $v_{AC}^{[t]}$ estimated on the *VM II* corpus. Left: voiced plosive sound /g/ and its unvoiced counterpart /k/. Right: voiced fricative sound /v/ and its unvoiced counterpart /f/.

4.5.3 Average Magnitude Difference

Motivation of average magnitude difference (AMD) is to measure periodicity in a time frame by summing up absolute differences between equidistant samples (see chapter "Pitch Detection" in [Rabiner and Schafer 1979]). Periodic signals yield zero values for distances of integer multiples of the period length and values larger than zero for all other distances. Assume the average magnitude difference $D^{[t]}(\tau)$ for some time frame t and a shift τ:

$$D^{[t]}(\tau) = \frac{1}{W-\tau} \sum_{\nu=0}^{W-\tau-1} |x^{[t]}(\nu) - x^{[t]}(\nu+\tau)|. \tag{4.38}$$

where W is the length of a time frame. Fig. 4.19 shows the average magnitude difference functions of two time frames extracted from the same utterance of the sound /a/. On the left hand side, equidistant peaks of values relative close to zero indicate that the time frame is voiced. The average magnitude difference depicted on the right hand side has been calculated for a time frame at the end of the utterance. The voicing is decaying to the end resulting in a less obvious peaks.

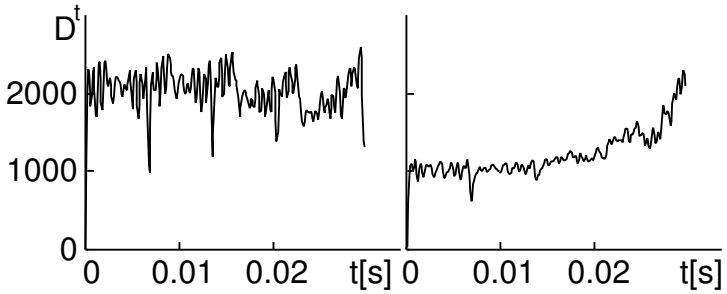


Figure 4.19: Average magnitude difference of two 40ms long time frames from the middle (left) and from the end (right) of an utterance of the sound /a/.

The definition of average magnitude difference resembles the one of the unbiased estimate of the autocorrelation. In the following, the connection of the two definitions is discussed. The average magnitude difference can be seen as a functional alternative of the autocorrelation. It measures the similarity of a time frame $x^{[t]}(\nu)$ of length W and its

time shifted copy $x^{[t]}(\nu+\tau)$ by applying absolute difference instead of multiplication. The functional correspondence can be shown by substituting the quadratic difference instead of the absolute one into Eq. (4.38):

$$Q^{[t]}(\tau) = \frac{1}{W-\tau}\sum_{\nu=0}^{W-\tau-1}|x^{[t]}(\nu)-x^{[t]}(\nu+\tau)|^2 \tag{4.39}$$

$$Q^{[t]}(\tau) = \frac{1}{W-\tau}\left(\sum_{\nu=0}^{W-\tau-1}x^{[t]}(\nu)^2+\sum_{\nu=\tau}^{W-1}x^{[t]}(\nu)^2\right)-2\tilde{R}^{[t]}(\tau) \tag{4.40}$$

$$Q^{[t]}(\tau) \approx 2(\tilde{R}^{[t]}(0)-\tilde{R}^{[t]}(\tau)). \tag{4.41}$$

As Eq. (4.40) shows, the average quadratic difference $Q^{[t]}(\tau)$ can almost be reduced to the unbiased estimate of the autocorrelation $\tilde{R}^{[t]}(\tau)$. The major difference can be found in the first term of Eq. (4.40). This term estimates the sum of the average energy of the original and the shifted signals. The average is calculated over the overlapping parts. $\tilde{R}^{[t]}(0)$ expresses the average energy as well but calculated over the complete time frame. Hence, the first term of Eq. (4.40) is equal to $2 * \tilde{R}^{[t]}(0)$ for $\tau = 0$ but it may differ from it at increasing values of τ.

In order to produce a measure of voicedness which is independent of the loudness of the frame, an upper bound of the average magnitude difference will be derived in the following. The upper bound can be found by applying the vector norm inequality $\|v\|_1 \leq \sqrt{n}\|v\|_2$ where n is the length of vector v. If we replace the elements of v by the difference $v^{[t]}(\nu) - v^{[t]}(\nu+\tau)$ and n by $W-\tau$ we get the following inequality:

$$0 \leq D^{[t]}(\tau) \leq \sqrt{Q^{[t]}(\tau)}. \tag{4.42}$$

Furthermore, a practically sufficiently high upper bound of $Q^{[t]}(\tau)$ can be found by applying the relation between the average quadratic difference and the unbiased estimate of the autocorrelation, see Eq. (4.41). Inserting the likewise practically sufficient upper bound $\tilde{R}^{[t]}(0) \geq |\tilde{R}^{[t]}(\tau)|$ into Eq. (4.41), the following upper bound is obtained:

$$D^{[t]}(\tau) \leq 2\sqrt{\tilde{R}^{[t]}(0)}. \tag{4.43}$$

The multiplicative factor 2 can be neglected since it does not carry any classification information. Hence, dividing the average magnitude difference by $\sqrt{\tilde{R}^{[t]}(0)}$ ensures in practice values in the range of [0..2]. The voicing measure $v_{AMD}^{[t]}$ can be defined as the minimum value of the normalized average magnitude difference in the interval of practically relevant pitch periods [2.5ms..12.5ms]:

$$v_{AMD}^{[t]} = \frac{\min\limits_{2.5\text{ms}\cdot f_s \leq t \leq 12.5\text{ms}\cdot f_s} D^{[t]}(t)}{\sqrt{\tilde{R}^{[t]}(0)}} \tag{4.44}$$

A value v_{AMD} close to 0 indicates periodicity thus voicing while greater values indicate less voicing in the time frames. Fig. 4.20 summarizes the necessary steps to extract the

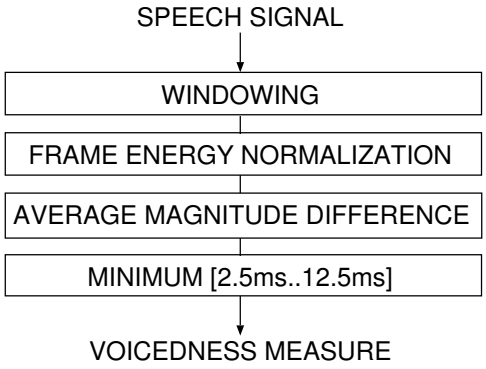


Figure 4.20: Block diagram of the average magnitude spectrum based voicing measure.

voicing measure. The first two step are identical to the extraction of the autocorrelation based measure. The frame energy normalization ensures here as well that $\tilde{R}^{[t]}(0) \equiv 1$ thus the division in Eq. (4.44) can be omitted. After calculating the average magnitude difference function for the discrete lags $[2.5\text{ms} \cdot f_s \; .. \; 12.5\text{ms} \cdot f_s]$, a simple linear search is used to find the minimal value. In this way, the one-dimensional average magnitude difference based voicing feature is generated every 10ms.

Finally, Fig. 4.21 shows histograms of $v_{AMD}^{[t]}$ on voiced-unvoiced sound pairs. The histograms have been estimated in the same way as described in Section 4.5.1. For the sake of easier comparability with Fig. 4.14 and Fig. 4.18, the histograms of $1 - v_{AMD}^{[t]}$ have been estimated. The average magnitude difference based voicing feature seems to separate voiced-unvoiced sounds similar to the autocorrelation based one. Recognition results presented in Section 4.5.4 have confirmed this observation.

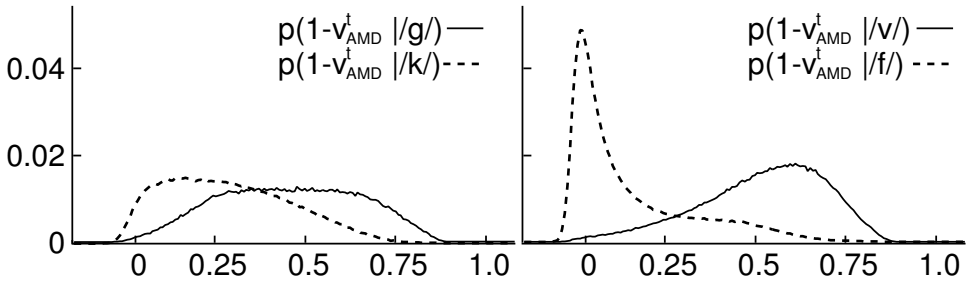


Figure 4.21: Histograms of measure $v_{AMD}^{[t]}$ estimated on the *VM II* corpus. For easier comparability with Fig. 4.14 and Fig. 4.18, values of $1 - v_{AMD}^{[t]}$ have been used instead of the original ones. Left: voiced plosive sound /g/ and its unvoiced counterpart /k/. Right: voiced fricative sound /v/ and its unvoiced counterpart /f/.

4.5.4 Comparison of Voicing Measures

After the introduction of three different extraction methods, a systematic experimental and theoretical comparison is presented in this section. The results presented in this section were first published in [Zolnay et al. 2003]. The comparison is started by the

description of experiments and their results in Section 4.5.4.1. A theoretical analysis follows the recognition results. The autocorrelation method is compared to the harmonic product spectrum in Section 4.5.4.2 and to the average magnitude difference based method in Section 4.5.4.3. This section is closed by the summary of the results in Section 4.5.4.4.

4.5.4.1 Experimental Results

A one-dimensional voicing measure has to be viewed as an auxiliary feature in contrast to the baseline features (e.g. MFCC). Therefore, the use of a voicing measure necessarily implies feature combination. Experiments have been performed on small- and large-vocabulary corpora. In the following experiments, the different voicing measures have been tested in combination with the MFCC feature. Feature combination has been carried out by using the LDA based method as descibed in Chapter 5. In Table 4.2, experimental results are summarized for the baseline MFCC system and for its combination with the different voicing measures. On the small-vocabulary *SieTill* corpora, a relative

Table 4.2: Word error rate on the *SieTill* and *VM II* corpora obtained by the combination of the MFCC feature with one of the voicing measures: harmonic product spectrum (HPS), autocorrelation (AC), or average magnitude difference (AMD). (Note that a within-word acoustic model has been used here for the *VM II* corpus.)

corpus	acoustic feature	error rates [%]		
		del	ins	WER
SieTill	MFCC	0.30	0.52	1.91
	MFCC+HPS	0.29	0.37	1.65
	MFCC+AC	0.28	0.40	1.71
	MFCC+AMD	0.28	0.38	1.71
VM II	MFCC	4.9	3.4	23.5
	MFCC+HPS	5.7	2.8	22.5
	MFCC+AC	5.7	2.9	22.8
	MFCC+AMD	5.7	2.8	22.2

improvement in WER of up to 14% has been obtained relative to using the MFCCs alone. The relative improvement on the large-vocabulary corpus are smaller. Improvements of up to 6% have been obtained relative to the extensively optimized baseline system. As shown in Table 4.2, the different extraction methods have performed on average similarly with small and inconsistent differences on the two corpora. Finally, experiments with combination of more than one voicing measure have not shown any improvement over using only one additional voicing measure.

4.5.4.2 Analytical Relation of AC and HPS

Recognition experiments did not show any significant difference between the autocorrelation (AC) and the harmonic product spectrum (HPS) based voicing measures. In this

section, the two methods are compared on a theoretical level, to verify the experimental results. For the sake of simplicity, the analysis is carried out in the continuous domain.

Let $x(t)$ be a signal in the time domain and $X(\jmath\omega) = \mathcal{F}\{x(t)\}$ its Fourier transform. Furthermore, the harmonic product spectrum $P(\jmath\omega)$ is defined as the product of copies of the power spectrum $|X(\jmath\omega)|^2$ shrunken by a factor of r:

$$P(\jmath\omega) = \prod_{r=1}^{R} |X(\jmath\omega r)|^2. \tag{4.45}$$

The autocorrelation is the integral of the product of the time domain signal $x(\nu)$ and its copy $x(\nu + \tau)$ shifted by a lag τ:

$$R_{xx}(\tau) = \int_{\nu=-\infty}^{\infty} x(\nu)\ x(\nu + \tau). \tag{4.46}$$

The relation of the two functions can be found by applying the inverse Fourier transform to the harmonic product spectrum. After the systematic application of the following lemmas:

$$\mathcal{F}^{-1}\{|X(\jmath\omega|^2\} = R_{xx}(\tau), \tag{4.47}$$

$$\mathcal{F}^{-1}\{X(\jmath\omega)Y(\jmath\omega)^*\} = R_{xy}(\tau) = \mathrm{corr}(x(t), y(t)), \tag{4.48}$$

$$\mathcal{F}^{-1}\{X(\jmath\omega r)\} = \frac{1}{r}x(\frac{t}{r}), \tag{4.49}$$

the inverse Fourier transform of $P(\jmath\omega)$ turns out to be the cross-correlation of the copies of the autocorrelation function $R_{xx}(\tau)$ stretched by a factors $r = 1, .., R$:

$$\mathcal{F}^{-1}\{P(\jmath\omega)\} = \tfrac{1}{r!}\mathrm{corr}(R_{xx}\left(\tfrac{t}{1}\right), \mathrm{corr}(R_{xx}\left(\tfrac{t}{2}\right), ...\mathrm{corr}(R_{xx}\left(\tfrac{t}{R-1}\right), R_{xx}\left(\tfrac{t}{R}\right))...)). \tag{4.50}$$

Analyzing the harmonic product spectrum in the time domain shows that whereas the autocorrelation based measure relies on the unchanged autocorrelation function, the harmonic product spectrum based method needs the cross-correlation of stretched copies of the autocorrelation function. Note that in case of $r = 1$, the two methods share the same underlying function. The next steps of the comparison are based on this simple case. In the case of $r = 1$, the harmonic product spectrum is equal to the power spectrum. If $x(t)$ is periodic, the power spectrum is discrete and the simple search for the maximum finds the harmonic with the highest magnitude. The interpretation of this maximum is not obvious. It is not enough to normalize the maximum by the frame energy to get a voicing measure since unvoiced frames can yield the same maximal value as well. The empirical solution proposed in this work divides the maximum by the average energy of its direct neighborhood rather than by the energy of the complete frame. The autocorrelation based methods do not only use a simple underlying function but the interpretation of the maximal value is straightforward as well. The maximal value can simply be divided by the frame energy yielding a voicing measure which is independent of the loudness of the uttered phoneme. Unvoiced frames have generally lower maxima since only periodic functions attain the maximum at lags other than $\tau = 0$.

Consequently, the autocorrelation based measure seems to be more practical. It performs similar to the harmonic product spectrum based measure and the extraction of the measure relies on a simpler and clearer algorithm.

4.5.4.3 Analytical Relation of AC and AMD

The relation of the autocorrelation (AC) based and the average magnitude difference (AMD) based voicing measures has already been derived in Section 4.5.3. It has been shown that the two methods differ only by the functional form of the similarity function which the signal $x^{[t]}(\nu)$ and its shifted copy $x^{[t]}(\nu+\tau)$ is compared with. The extraction of the voicing measures share exactly the same normalization step and both use a simple search to find the maximum in case of the autocorrelation and the minimum in case of the average magnitude difference.

Autocorrelation uses multiplication as similarity function. It has been shown in Eq. (4.40), that using the multiplication is mainly identical to using the quadratic difference as similarity function. Only difference is in the average energy term. Consequently, the autocorrelation and the average magnitude difference functions can be merged into a generalized form:

$$G^{[t]}(\tau) = \frac{1}{W-\tau} \sum_{\nu=0}^{W-\tau-1} |x^{[t]}(\nu) - x^{[t]}(\nu+\tau)|^p. \tag{4.51}$$

Despite the better understanding of the relation of the autocorrelation and the average magnitude spectrum, the generalized form has not yielded any practical advantage. Experiments performed by using $p > 2$ have not resulted in any improvements in WER.

The two voicing measures are highly related. Nevertheless, the autocorrelation based method has a significant advantage. The calculation of the average magnitude spectrum has a complexity of $\mathcal{O}(W^2)$ when calculated for the complete frame of length W. The autocorrelation corresponds to the power magnitude spectrum in the frequency domain. This calculation of the power spectrum has a linear complexity in the frame length. The transformation of the time frames into the frequency domain can be carried out by using the Fast Fourier Transform (FFT) which has a complexity of $\mathcal{O}(W \log W)$. Along with the inverse FFT, the complexity of the autocorrelation calculation is $\mathcal{O}(W + 2W \log W)$ which is a much lower $\mathcal{O}(W^2)$ for window length of $W \approx 40\text{ms} * 16\text{kHz} = 640$.

Consequently, the autocorrelation based measure seems to be more practical in this case as well. It performs similar to the average magnitude difference based measure and the extraction of this measure can be carried out significantly faster by using the FFT algorithm.

4.5.4.4 Summary

In the Section 4.5.4, experimental and theoretical comparisons of three voicing measures have been presented. The autocorrelation based voicing measure has turned out to be the most practical according to the following results:

- Experiments have not shown any significant difference between the three methods neither on the small- nor on the large-vocabulary task.
- It has a simple and clear motivation and implementation.

- It can be implemented using FFT which leads to faster computation of the voicing measure.

In the next Chapters, feature combination experiments have been carried out using only the autocorrelation base voicing extraction. Consequently, the autocorrelation measure is often referred to as *voicing measure* and instead of the abbreviation AC the simple V is used.

4.5.5 Optimization of Autocorrelation Method

The empirical optimization of two aspects of the autocorrelation based voicing measure is presented in this Section. The first group of experiments presented in Section 4.5.5.1 has been performed in order to study the effects of time frame length on the voicing measure and consequently on the WER. The second group of experiments has investigated the estimation of the autocorrelation function. In Section 4.5.5.2, the unbiased estimate is compared to a method using the cross-correlation of time frames of different lengths. The experiments on the autocorrelation based method have been repeated by using the average magnitude difference to cross-validate the results with a slightly different method.

4.5.5.1 Frame Length

Before analyzing the experimental results, lower and upper bounds of the optimal frame length are derived. A lower bound for the frame length can be derived from the lowest excepted fundamental frequency ($\approx$ 80Hz). To detect periodicity, at least two periods of the signal have to fit into a time frame. Thus, in order to detect voiced segments of a fundamental frequency of 80Hz, a minimal frame length of $2 \cdot 1/80\text{Hz} = 25\text{ms}$ is required. An upper bound of the optimal frame length can be derived based on the average phoneme length ($\approx$ 60ms). Time frames containing two or more phonemes at a time cannot be used to analyze single voiced phonemes. The experiments have been carried out in combination with the MFCC feature. Table 4.3 summarizes the results obtained by using increasing window length for the voicing feature. The window length of the MFCC feature has been kept unchanged. Experiments performed by using the autocorrelation method resulted in an optimal window length of 40ms. The average magnitude difference yielded the best WER at a window length of 30ms. Nevertheless, a window length of 40ms has been used in the rest of the voicing experiments. The optimal value of 40ms corresponds to former results, cf. chapter Pitch Detection in [Rabiner and Schafer 1979].

4.5.5.2 Estimation of Autocorrelation Function

In Section 4.5.2, the autocorrelation function has been calculated by using the unbiased estimate. The unbiased estimate accounts for the decreasing amount of overlap between the windowed speech signal $x^{[t]}(\nu)$ and its shifted copy $x^{[t]}(\nu+\tau)$ when increasing τ. E.g., for the greatest value of $\tau = W - 1$, the autocorrelation function is estimated by the product of a single sample pair. In order to ensure constant amount of summands in the estimation of the autocorrelation, a windowing scheme depicted on Fig. 4.22 has been

Table 4.3: Word error rate on the *VM II* corpus obtained by increasing the window length W of the voicing measure. The tests have been carried out by combining the MFCC feature with two different voicedness measures: autocorrelation, average magnitude difference. (Note that the within-word acoustic model has been used here.)

voicedness measure	W [ms]	error rates [%] del	ins	WER
autocorrelation	20	5.9	3.0	23.6
	30	5.7	3.0	23.2
	40	5.8	2.9	22.6
	50	5.7	2.9	22.8
avr. magnitude difference	20	5.6	3.0	23.0
	30	5.6	2.6	22.2
	40	5.5	2.7	22.5
	50	5.6	2.8	22.5

tested. In this scheme the shifted copy of the time frame $x^{[t]}(\nu+\tau)$ is replaced by a time

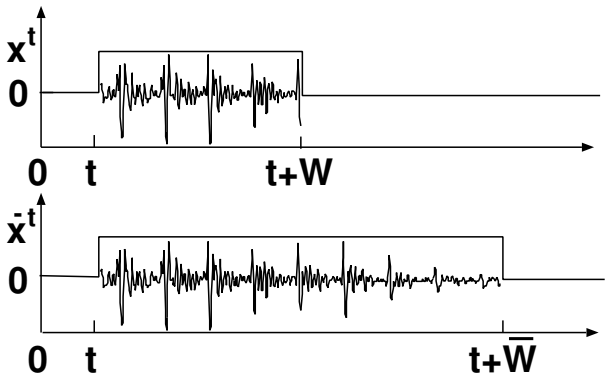


Figure 4.22: Windowing scheme for the cross-correlation based voicing measure.

frame of double length $\bar{x}^{[t]}(\nu+\tau)$.

$$\bar{R}^{[t]}(\tau) = \sum_{\nu=0}^{W-1} x^{[t]}(\nu)\ \bar{x}^{[t]}(\nu+\tau). \tag{4.52}$$

A double frame length of the shifted signal ($\bar{W} = 2W$) ensures that the length of the overlap remains constant. In Section 4.5.2, it has been noted that the unbiased estimate does not share theoretically all the properties of the autocorrelation function. One of the main differences is that the maximal value is not necessarily attained at $\tau = 0$. The same holds for this windowing scheme. Strictly speaking, $\bar{R}^{[t]}(\tau)$ is the cross-correlation of two different signals. The resulting function is neither necessarily symmetric nor does it necessarily attain its maximum at $\tau = 0$. Nevertheless, since $x^{[t]}(\nu)$ and $\bar{x}^{[t]}(\nu)$ are equal for $\nu < W$, the cross-correlation function is in practice a good estimate of the autocorrelation and it mostly shares the properties of the autocorrelation function.

The extraction of the voicedness measure requires an upper bound for the cross-correlation function. An upper bound analog to the ones presented formerly has been chosen. Applying the Cauchy-Schwartz Inequality to Eq. (4.52), the following upper bound is obtained:

$$\bar{R}^{[t]}(\tau) <= \sqrt{\sum_{\nu=0}^{W-1} x^{[t]}(\nu)^2 \sum_{\nu=0}^{W-1} \bar{x}^{[t]}(\nu)^2} = \sqrt{E^{[t]}\bar{E}^{[t]}}. \tag{4.53}$$

Using the upper bound, a voicing measure based the cross-correlation function Eq. (4.52) can be derived as follows:

$$v_{CC}^{[t]} = \frac{\max\limits_{2.5\text{ms}\cdot f_s \leq \tau \leq 12.5\text{ms}\cdot f_s} \bar{R}^{[t]}(\tau)}{\sqrt{E^{[t]}\bar{E}^{[t]}}}. \tag{4.54}$$

where f_s denotes the sample rate. Fig. 4.23 summarizes the necessary steps to extract the voicing measure. The extraction of the cross-correlation based voicing feature is a

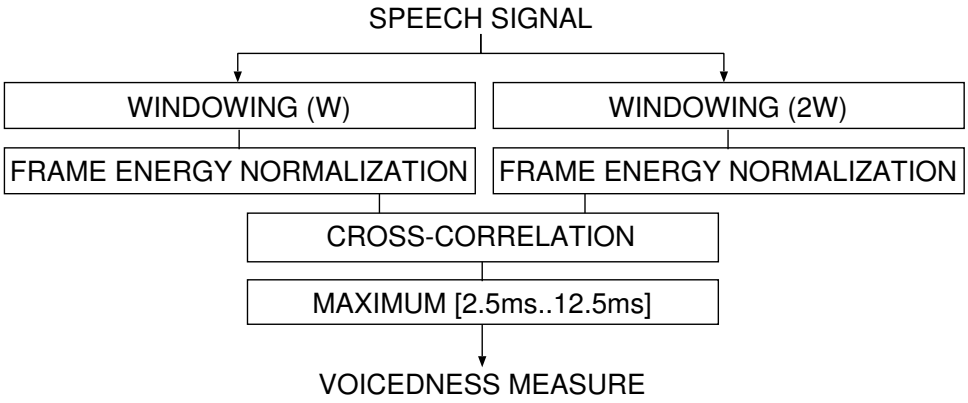


Figure 4.23: Block diagram of the cross-correlation based voicing measure.

generalization of the autocorrelation based one. The only difference is that two frames are processed parallel instead of one. The left windowing node outputs every 10ms a time frame of 40ms length whereas the right one a time frame of the length of 80ms. The frame energy normalization ensures that $E^{[t]} \equiv 1$ and $\bar{E}^{[t]} \equiv 1$. Hence the division in Eq. (4.54) can be omitted. After calculating the cross-correlation for the discrete lags $[2.5\text{ms} \cdot f_s \;..\; 12.5\text{ms} \cdot f_s]$, a simple linear search is used to find the maximal value. In this way, an one-dimensional voicing feature is generated every 10ms based on the cross-correlation of time frames of different lengths.

Table 4.4 summarizes recognition results obtained by using the unbiased estimate of the autocorrelation and the cross-correlation. The cross-correlation scheme can be transferred straightforward to the average magnitude difference based measure. Neither of the cross-correlation measures have yielded significant improvement compared to the autocorrelation based measures. Because of the lower computational costs, the unbiased estimate of the autocorrelation has been chosen for the feature combination experiments in this work.

Table 4.4: Word error rate on the *VM II* corpus obtained by using auto- and cross-correlation while calculating the voicing measures. The tests have been carried out by combining the MFCC feature with two different voicedness measures: autocorrelation, average magnitude difference. (Note that the within-word acoustic model has been used here.)

voicedness feature	$\bar{W}$	error rates [%]		
	$W = 40$ms	del	ins	WER
autocorrelation	W	5.8	2.9	22.6
cross-correlation	$2W$	5.4	2.8	22.5
average magnitude diff.	W	5.5	2.7	22.5
cross average magnitude diff.	$2W$	5.4	2.9	22.4

4.5.6 Recognition Results

In the previous sections, different aspects of the voicing feature extraction have been presented. The voicing measure based on the unbiased estimate of the autocorrelation function (V) has been shown to be the most practical without significant loss in WER. Table 4.5 summarizes recognition results obtained by using different setups. Recognition tests have been carried out in combination with different baseline features. Beside the MFCC feature, the MF-PLP and the PLP features have been tested as well. The voicing feature has also been tested in combination with the speaker adapted VTLN feature. The LDA based feature combination method has been used as described in Chapter 5. The experiments have been performed on the small-vocabulary *SieTill* corpus and on the large-vocabulary *VM II* and *EPPS* corpora. The application of the voicing measure has resulted in consistent improvements in WER over the different baseline features and corpora. The experiments yielded relative improvement of 11% on the small- and 3% on the large-vocabulary corpora relative to the baseline features.

An analysis of the recognition results is presented in the following by means of histograms estimated for the voiced fricative sound /v/ and its unvoiced counterpart /f/. The histograms have been estimated in the same way as described in Section 4.5.1. On Fig. 4.18, histograms of the voicing measure can be seen estimated for the same phoneme pair. The histograms show that the voicing measure contains significant amount of information which can help to distinguish the voiced-unvoiced phoneme pair /v/-/f/. Histograms shown on Fig. 4.24 have been estimated to measure how far the voicing measure can influence the recognition results when combined with the MFCC feature. The histograms have been estimated first using solely the MFCC feature and then using the LDA based combination of the MFCC and voicing features. Since both experiments applies LDA in a similar way, the histograms have been estimated on first three LDA output coefficients belonging to the three largest eigenvalues. Note that according to the definition of LDA, the larger an eigenvalue is the more classification information it contains with respect to the underlying classes. Although the histograms on the left hand side of Fig. 4.24 have been estimated to be the reference, one interesting fact can already be stated. Only the very first y_0 out of 45 LDA output coefficients yields a relatively small overlap of

Table 4.5: Word error rates obtained by the LDA based combination of baseline features (MFCC, VTLN, MF-PLP, PLP) and the voicing feature (V).

corpus	baseline feature	V	error rates [%] del	ins	WER
SieTill	MFCC	no	0.3	0.5	1.8
		yes	0.3	0.4	1.6
VM II	MFCC	no	4.5	2.9	21.0
		yes	4.6	2.7	20.3
	VTLN	no	3.8	2.9	19.1
		yes	4.1	2.7	18.7
	MF-PLP	no	5.2	2.3	21.0
		yes	4.7	2.6	20.5
	PLP	no	5.9	2.3	21.4
		yes	4.6	3.0	20.6
EPPS	MFCC	no	4.3/3.8	1.4/1.7	14.7/15.3
		yes	3.9/3.4	1.5/1.9	14.3/14.8
	VTLN	no	4.3/3.7	1.3/1.5	14.2/14.1
		yes	4.0/3.3	1.5/1.6	13.8/14.0
	MF-PLP	no	4.2/3.7	1.5/1.7	14.8/15.3
		yes	3.7/3.2	1.7/2.0	14.3/15.2
	PLP	no	4.3/3.5	1.6/1.8	15.4/15.8
		yes	4.5/3.7	1.4/1.6	15.1/15.4

histograms when using solely the MFCC feature. This indicates that out of 45 coefficients, only a single one contributes to differentiating the phoneme pair /v/-/f/. The overlap of different histograms has been collected in Table 4.6. When using the additional voicing

Table 4.6: Overlap of histograms shown on Fig. 4.24.

LDA output index	overlap of histograms: /v/ and /f/ [%] LDA(MFCC) (y)	LDA(MFCC+V) ($\tilde{y}$)
0	24	18
1	69	71
2	62	59

measure, histograms estimated for the first LDA output $\tilde{y}_0$ indicate that the discrimination power of the combined features has been improved. The overlap of histograms has been reduced from 24% to 18%. The rest of the 45 coefficiets have not been improved by the additional voicing feature.

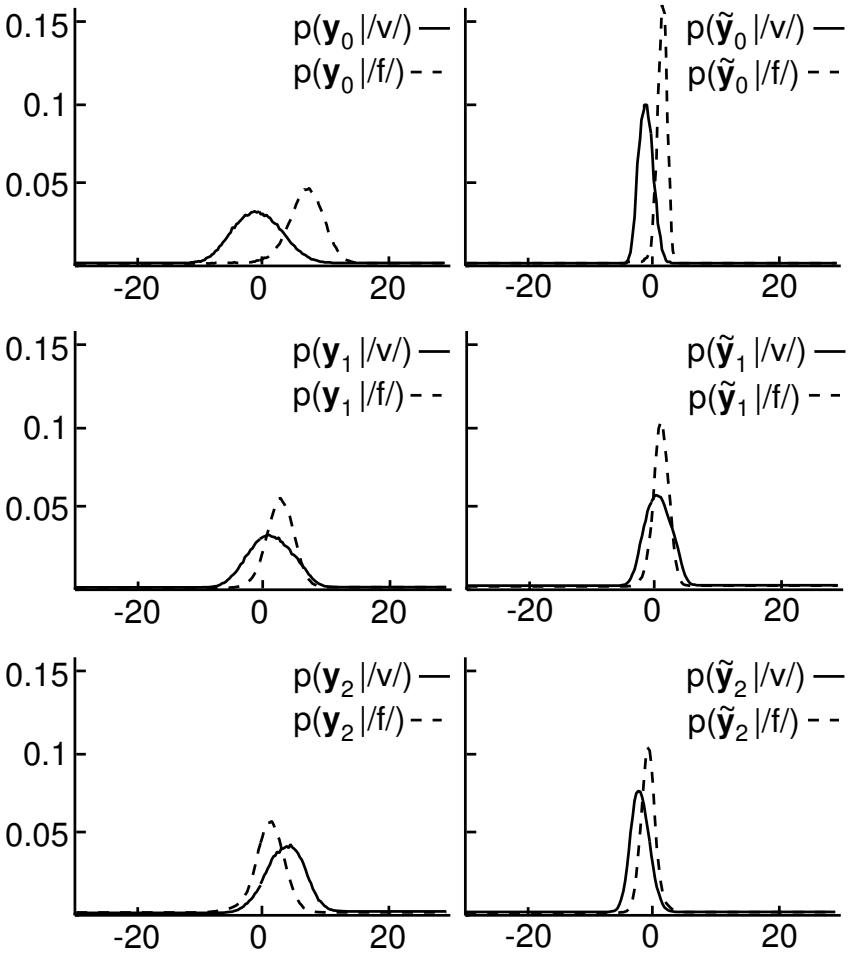


Figure 4.24: Histograms of LDA output coefficients belonging to the three largest eigenvalues. Histograms have been estimated on the *VM II* corpus for the voiced fricative sound /v/ and its unvoiced counterpart /f/. Left: first three output coefficients (y_0, y_1, y_2) of LDA applied solely to the MFCC feature. Right: first three output coefficients $(\tilde{y}_0, \tilde{y}_1, \tilde{y}_2)$ of the LDA based combination of the MFCC and voicing features.

4.5.7 Summary

In this Section, three extraction methods have been presented for the voicing feature. Although, these methods share a common motivation and consequently a common structure, there are significant differences in their implementations. Recognition tests performed on small- and large-vocabulary tasks have not shown any difference between the three methods. A theoretical comparison has shown that the autocorrelation based voicing measure is the best choice for later feature combination experiments due to its clear motivation and its low computational cost. After the optimization of the parameters of the autocorrelation based measure, comprehensive recognition tests have been presented. The measure has been tested on three different corpora and in combination with four different baseline features among others a speaker adapted feature. Results have shown that using the additional voicing measure yields consistent improvements in WER compared to using

the single baseline features.

4.6 Spectrum Derivative Feature

The spectrum derivative feature (SD) has been developed in cooperation with Daniil Kocharov from the Department of Phonetics, Saint-Petersburg State University. The feature focuses on the differentiation of two articulatory classes: obstruent and sonant consonants. Obstruent consonants are formed by obstructing the airway causing turbulence in the airflow whereas voiced consonants are called sonant consonants. From a phonetic point of view, these two classes differ by the presence of formants. In the magnitude spectrum of sonants, peaky formant-like structures can be observed. However obstruents manifest in a flat and noisy magnitude spectrum. Thus, a feature summarizing the intensity of changes of the magnitude spectrum over the frequency axis can help to differentiate these two phoneme classes. The first results obtained by using the spectrum derivative have been published in [Kocharov et al. 2005].

In Section 4.6.1, a detailed description of the extraction algorithm of the spectrum derivative feature is presented followed by the optimization of parameters in Section 4.6.2. In Section 4.6.3, an analysis of histograms of the feature is carried out over different phoneme classes and recognition results are presented obtained on small- and large-vocabulary corpora.

4.6.1 Extraction Algorithm

The spectrum derivative feature is a measure calculated as the absolute sum of the first order derivative of the magnitude spectrum. The extraction procedure is shown on Fig. 4.25. A Hamming window is applied to preemphasized speech segments. The frame shift

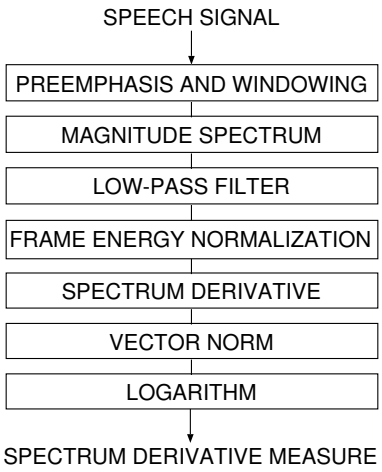


Figure 4.25: Block diagram of spectrum derivative measure.

is chosen to 10ms. The frame length has been optimized empirically in a range from 15ms

to 90ms. Experiments presented in Section 4.6.2 have shown that the best results can be obtained by using 25ms. The magnitude spectrum $X^{[t]}[n]$ of time frame t is calculated by using FFT along with an appropriate zero padding. The preprocessing of the magnitude spectrum begins with discarding the high frequency magnitudes. The cut-off frequency is chosen at 1kHz. The implementation applies an ideal low-pass filter:

$$\hat{X}^{[t]}[n] = \begin{cases} X^{[t]}[n] & n < \frac{1\text{kHz} \cdot N}{f_s} \\ 0 & \text{otherwise} \end{cases} \tag{4.55}$$

where N denotes the number of FFT points and f_s the sample rate. Comparative tests have shown on different corpora that processing only the lower part of the magnitude spectrum yields the best recognition results. (The test are described in Section 4.6.2.) In the next step, the filtered magnitude spectrum $\hat{X}^{[t]}[n]$ is energy normalized to account for different frame energies. Experiments have been carried out by using frame-wise and utterance-wise energy normalization. Best recognition results have been obtained by applying energy normalization to every time frame:

$$\tilde{X}^{[t]}[n] = \frac{\hat{X}^{[t]}[n]}{\sqrt{\hat{X}^{[t]}[0]^2 + \hat{X}^{[t]}[\frac{N}{2}]^2 + 2\sum_{n=1}^{N/2-1} \hat{X}^{[t]}[n]^2}}, \tag{4.56}$$

where t denotes the time frame, n denotes the discrete frequency, and N is the number of FFT points. The first order derivative $a^{[t]}[n]$ is calculated over the normalized magnitude spectrum $\tilde{X}[n]$:

$$a^{[t]}[n] = \tilde{X}^{[t]}[n] - \tilde{X}^{[t]}[n-1], \tag{4.57}$$

$$a^{[t]}[0] \equiv 0. \tag{4.58}$$

Finally, the spectrum derivative feature is a continuous measure $s^{[t]}$ calculated as the logarithm of the absolute sum of the first order derivative:

$$s^{[t]} = \log\left(\sum_{n=0}^{N/2} |a^{[t]}[n]|\right). \tag{4.59}$$

Note that this method can be straightforward extended by using higher order derivatives. The measure can be calculated for every higher order derivative of the magnitude spectrum as well. Nevertheless, the experiments presented in Section 4.6.2 have not shown any additional improvement in WER when using the higher order derivatives on top of the first order derivative. Therefore, feature combination experiments including the spectrum derivative feature use solely the first order derivative producing a single measure.

4.6.2 Optimization of Spectrum Derivative Measure

In this section, three major parameters of the spectrum derivative feature are analyzed in terms of WER. The parameters are optimized empirically on different corpora. Beside the window length of the feature, the cut-off frequency of the low-pass filter and the usage of higher order derivatives are studied in the following.

The experiments optimizing the window length have been carried out on the *SieTill* corpus. Similar to the voicing feature, also the spectrum derivative feature has to be viewed as auxiliary. Therefore, recognition results were also produced by using the LDA based feature combination. Window lengths have been tested in the interval [15ms..90ms]. Experiments summarized in Table 4.7 have shown that the best results can be obtained by using 25ms which is the same frame length used when generating the MFCC feature.

Table 4.7: Word error rates obtained by the LDA based combination of the MFCC and the spectrum derivative (SD) features. The window length of the spectrum derivative measure has been increased from 15ms to 90ms. Experiments have been performed on the *SieTill* corpus.

SD window length [ms]	error rates [%] del	ins	WER
0 (MFCC)	0.50	0.71	3.83
15	0.48	0.55	3.35
25	0.43	0.67	3.32
40	0.46	0.55	3.35
50	0.49	0.73	3.39
60	0.49	0.80	3.46
70	0.51	0.80	3.51
80	0.52	0.88	3.66
90	0.52	0.94	3.78

The introduction of low-pass filtering is due to promising trial-and-error experimental results. Low-pass filtering is one of the ideas which have yielded improvements in the WER. The tests have been carried out on two large-vocabulary tasks. The first order spectrum derivative measure has been tested in combination with the baseline MFCC feature. Table 4.8 presents the recognition results obtained by increasing the cut-off frequency of the ideal low-pass filter from 500Hz to 6000Hz. On the *VM II* corpus, the spectrum derivative feature yields an improvement in WER similar to the voicing feature when using a cut-off frequency of 1kHz. With increasing cut-off frequencies, the 3% relative improvement in WER slowly disappears. Recognition tests have not yielded any consistent improvement on the *EPPS* corpus. Hence, a cross-validation experiment has been performed using the optimal cut-off frequency, 1kHz, on the *SieTill* corpus. The improvement caused by the additional spectrum derivative measure has not reduced when applying the low-pass filter. Nevertheless, further experiments are necessary to understand the effects of the low-pass filtering. In feature combination experiments applying the spectrum derivative feature, the optimal cut-off frequency of 1kHz has been used.

Finally, experiments are presented using spectrum derivative measures derived not only from the first but also from higher order derivatives. The ith order derivatives $a_t^{(i)}[n]$ of the preprocessed magnitude spectrum $\tilde{X}^{[t]}[n-1]$ of the time frame t are calculated in

Table 4.8: Word error rates obtained by increasing the cut-off frequency of the low-pass filter. The spectrum derivative measure is tested in the LDA based combination with the MFCC feature.

corpus	cut-off frequency [Hz]	error rates [%] del	ins	WER
VM II	0 (MFCC)	4.5	2.9	21.0
	500	4.7	3.1	20.9
	1000	4.5	2.9	20.3
	2000	5.2	2.8	20.7
	4000	4.4	3.3	20.9
	6000	5.5	2.4	20.8
EPPS	0 (MFCC)	4.3/3.8	1.4/1.7	14.7/15.3
	500	4.1/3.6	1.5/1.8	14.6/15.3
	1000	4.5/4.0	1.2/1.5	14.7/15.1
	2000	4.3/3.7	1.4/1.7	14.8/15.0
	4000	4.2/3.8	1.3/1.6	14.5/15.1
	6000	4.3/3.6	1.4/1.7	14.8/15.3

the following way:

$$a_t^{(i)}[n] = a_t^{(i-1)}[n] - a_t^{(i-1)}[n-1], \tag{4.60}$$

$$a_t^{(0)}[n] = \tilde{X}^{[t]}[n] - \tilde{X}^{[t]}[n-1], \tag{4.61}$$

$$a_t^{(i)}[0] \equiv 0. \tag{4.62}$$

$$\tag{4.63}$$

Thus, the ith order derivatives $a_t^{(i)}[n]$ are calculated by reapplying Eq. (4.57) to the $(i-1)$st order derivatives. The actual spectrum derivative measures $s_t^{(i)}[n]$ are calculated applying Eq. (4.59) to the higher order derivatives:

$$s_t^{(i)}[n] = \log\left(\sum\nolimits_{n=0}^{N/2} |a_t^{(i)}[n]|\right). \tag{4.64}$$

Table 4.9 summarizes the recognition results obtained by using more than one spectrum derivative measures. Note when applying the ith order derivative, all the lower order derivatives $1, 2, ..., i-1$ are used as well. The spectrum derivative measures are tested in combination with the MFCC and the voicing (V) features. The LDA based feature combination is applied to find an optimal linear combination of the feature coefficients. On the small-vocabulary *SieTill* task, increasing the number of spectrum derivatives resulted in inconsistent and small changes in the WER compared to using solely the first order measure. On the large-vocabulary task *VM II*, no improvements have been obtained when adding measures derived from higher order derivatives. Therefore, higher order measures are abandoned in feature combination experiments including the spectrum derivative feature. Nevertheless, further investigations are required to study on the role of higher order measures.

Table 4.9: Word error rates obtained by increasing number of spectrum derivative measures. The spectrum derivative measures are tested in the LDA based combination with the MFCC and the voicing (V) features. Note when adding the ith order spectrum derivative, all the lower order $(1, 2, ..., i\text{-}1)$ ones are used as well.

corpus	#SD	error rates [%]		
		del	ins	WER
SieTill	0 (MFCC+V)	0.27	0.38	1.52
	1	0.26	0.34	1.51
	2	0.26	0.37	1.60
	3	0.24	0.33	1.45
	4	0.25	0.34	1.51
	5	0.24	0.35	1.53
VM II	0 (MFCC+V)	4.8	2.8	21.2
	1	4.6	2.8	20.8
	2	4.4	3.0	21.4
	3	5.1	2.8	21.3

4.6.3 Results and Analysis

In order to analyze the spectrum derivative feature, histograms of the first order spectrum derivative measure have been generated for a specific phoneme pair. Fig. 4.26 depicts distributions of $s^{[t]}$ on the exemplary phoneme pair /v/ and /s/, which, phonetically, differ by their sonority. Histogram of a given phoneme has been estimated on time frames aligned to any of the states of the triphones with the given phoneme as central one.

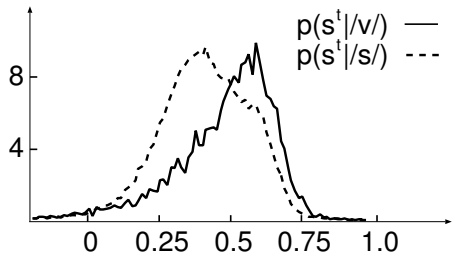


Figure 4.26: Histograms of spectrum derivative measure $s^{[t]}$ for the sonant consonant /v/ and the obstruent consonant /s/ estimated on the *VM II* corpus.

Finally, recognition tests have been performed on different corpora and in combination with different features to test the generality of the feature. Solely the first spectrum derivative measure (SD) has been used calculated by applying a window length of 25ms and a low-pass filter with a cut-off frequency of 1kHz. The spectrum derivative measure has been tested in combination with the MFCC and VTLN features on small- and large-vocabulary corpora. Features have been combined by using the LDA based method as described in Chapter 5. Table 4.10 shows the results obtained by using the single

additional spectrum derivative measure. Adding the spectrum derivative measure has

Table 4.10: Word error rates obtained by the LDAbased combination of baseline features (MFCC, VTLN) and the spectrum derivative feature (SD).

corpus	baseline feature	SD	error rates [%] del	ins	WER
SieTill	MFCC	no	0.3	0.5	1.8
		yes	0.3	0.4	1.6
VM II	MFCC	no	4.5	2.9	21.0
		yes	4.5	2.9	20.3
	VTLN	no	3.8	2.9	19.1
		yes	3.7	3.0	18.6
EPPS	MFCC	no	4.3/3.8	1.4/1.7	14.7/15.3
		yes	4.5/4.0	1.2/1.5	14.7/15.1
	VTLN	no	4.3/3.7	1.3/1.5	14.2/14.1
		yes	4.0/3.3	1.5/1.7	14.2/14.1

resulted in improvements in WER of 11% on the *SieTill* and 3% on the *VM II* corpora relative to the baseline features. The spectrum derivative feature could not improve the WER significantly on the *EPPS* corpus.

4.6.4 Summary

A recently developed acoustic feature, the spectrum derivative feature has been described in this section. Several aspects of the extraction of the measure have been analyzed and optimized. Recognition tests have been carried out using several different setups. The spectrum derivative measure shows a performance similar to the voicing feature when combined with state-of-the-art acoustic features. A relative improvement in WER of 11% has been obtained on a small-vocabulary task. The improvements on the large-vocabulary tasks have not been consistent. Whereas no significant changes have been observed on the *EPPS* corpus, the additional spectrum derivative measure has resulted in relative imporvements of up to 3% on the *VM II* corpus.

Chapter 5

LDA Based Feature Combination

The Linear Discriminant Analysis (LDA) is a linear transformation which projects a feature space onto a lower dimensional subspace such that the class separability for distributions with equal variances is maximized [Fisher 1936, Duda et al. 2001]. In statistical classification tasks, LDA is mainly used to reduce the dimension of the feature space with the goal of conveying the most relevant classification information into the reduced feature space. In speech recognition, the state-of-the-art feature extraction techniques produce $\approx$ 15 coefficients. Including first and second order derivatives of the coefficients, a feature space of $\approx$ 45 components is produced which is an optimal dimension for corpora consisting of $\approx$ 50 hours of training set. Hence, a direct application of LDA to each time frame is not beneficial. In [Häb-Umbach and Ney 1992], LDA was first applied successfully to speech recognition. Instead of applying LDA to single feature vectors, LDA has been used to find an optimal linear combination of successive feature vectors. For every time frame, $\approx$ 11 successive time frames have been concatenated and LDA has been used to reduce the concatenated large ($\approx$ 200 coefficients) feature space onto a subspace of $\approx$ 40 output coefficients. This method can be seen as a generalization of the dynamic features. An introduction to the RWTH implementation is provided in [Welling 1999].

The LDA based feature combination approach can be derived in a straightforward way from the aforementioned technique. In Section 5.1, a review of the theory of LDA is presented along with practical remarks on its applications. In Section 5.2, the procedure is presented in which LDA has been used for feature combination. In Section 5.3, important differences are pointed out between the setups on the small- and large-vocabulary corpora. Recognition results using the LDA based combination of different feature sets are presented in Section 5.4. Results are presented for using two sets of features. Combining the first set of features has resulted in a consistent improvement in WER when increasing the number of combined features. However, tests performed using the second set have yielded degradations in WER even compared to the worst performing model in the set. A study of this degradations is carried out in Section 5.5 along with the review of observations made by other research groups.

5.1 Theory of Linear Discriminant Analysis

Linear Discriminant Analysis [Fisher 1936, Duda et al. 2001] is a dimensionality reducing transformation. Dimensionality reducing transformations project a feature space of dimension $\mathcal{D}$ onto a lower dimensional one with dimension $D < \mathcal{D}$. The objective of LDA is twofold. On the one hand, it aims at reducing the dimensionality of the feature space in order to care for data sparseness problems that may occur in high dimensional feature spaces. On the other hand, it aims at retaining a maximum of class separability information. The transformation can be defined as follows:

$$\begin{aligned} F_{V_D} : \mathbb{R}^{\mathcal{D}} &\rightarrow \mathbb{R}^{D} \quad (D < \mathcal{D}) \\ x &\rightarrow y = V_D^T \tilde{x} \ , \end{aligned} \tag{5.1}$$

where V_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. The objectives of the estimation of the LDA matrix V_D are the following:

- the distances within one class remain constant;
- the distances between the class centers are maximized.

The training data is given in the form of feature vector x_n and class index k_n pairs for each observation $n = 1, \ldots, N$. In order to measure the within- and between-class distances of the transformed $V_D^T x_n$ observations, the following definitions are introduced:

Total mean vector:

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n, \tag{5.2}$$

Vectors belonging to a class k and their mean vector:

$$A_k = \{x_n : k = k_n\}, \tag{5.3}$$

$$\mu_k = \frac{1}{n_k} \sum_{x \in A_k} x, \tag{5.4}$$

$$\tag{5.5}$$

where $n_k = |A_k|$ denotes the number observation belonging to class k. The distances within and between classes are measured by the euclidean distance $|| \cdot ||^2$:

Sum of within-class distances:

$$D_W(V_D) = \sum_k \sum_{x \in A_k} ||V_D^T (x - \mu_k)||^2. \tag{5.6}$$

Sum of between-class distances:

$$D_B(V_D) = \sum_k n_k ||V_D^T (\mu_k - \mu)||^2. \tag{5.7}$$

Let $tr(X) := \sum_{d=1}^{\mathcal{D}} X_{dd}$ be the *trace* of a matrix $X \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ such that $u^T u = ||u||^2 = tr(uu^T)$ for $u \in \mathbb{R}^{\mathcal{D}}$. Applying trace to Eq. (5.6) and Eq. (5.7), we obtain the within- and between-class scatter matrices W and B, respectively:

$$
\begin{aligned}
D_W(V_D) &= tr\Big[\sum_k \sum_{x\in A_k} V_D^T (x-\mu_k)(x-\mu_k)^T V_D\Big] && (5.8)\\
&= tr\Big[V_D^T \underbrace{\sum_k \sum_{x\in A_k} (x-\mu_k)(x-\mu_k)^T}_{W} V_D\Big], && (5.9)\\
D_B(V_D) &= tr\Big[\sum_k n_k V_D^T (\mu_k-\mu)(\mu_k-\mu)^T V_D\Big] && (5.10)\\
&= tr\Big[V_D^T \underbrace{\sum_k n_k (\mu_k-\mu)(\mu_k-\mu)^T}_{B} V_D\Big]. && (5.11)\\
& && (5.12)
\end{aligned}
$$

Consequently, the optimization criterion $F(V_D)$ for the transformation matrix V_D can be defined as:

$$F(V_D) = \frac{tr[V_D^T B V_D]}{tr[V_D^T W V_D]}. \tag{5.13}$$

The maximization of this criterion is a well known problem in the linear algebra. Applying the *Generalized Eigenvalue Problem* delivers the optimal matrix V_D:

$$Bx = \lambda W x, \tag{5.14}$$

where λ denotes an eigenvalue and x an eigenvector. If exists, the generalized eigenvalue problem yields a complete system of eigenvectors, containing $\mathcal{D}$ optimal projection vectors. The eigenvalue indicates how much between-class variation falls onto the direction of the corresponding eigenvector. Hence, the optimal mapping V_D consists of the eigenvectors belonging to the first largest D eigenvalues. In order to motivate the use of the generalized eigenvalue problem, a direct derivation is provided focusing on finding a one-dimensional projection v of the feature vectors. In this case the trace can be neglected in the optimization criterion Eq. (5.13):

$$f(v) = \frac{v^T B v}{v^T W v}. \tag{5.15}$$

This simplified criterion can be interpreted as the ratio of two variances. The variances are measured only in the one-dimensional subspace given the projection v. The variance of the class centers falling on the direction v is given by the nominator whereas the variance within classes falling on the direction v can be found in the denominator. Since this criterion is invariant with respect to the length of the projector vector v, a constraint on the within-class variance can be introduced without the loss of generality:

$$v^T W v := 1. \tag{5.16}$$

Introducing a Lagrange multiplier for this constraint, the optimum criterion can be written as:

$$f(v, \lambda) = v^T B v + \lambda(v^T W v - 1). \tag{5.17}$$

After deriving the criterion with respect to v and *lambda*, the generalized eigenvalue problem is retrieved. This implies that the maximization of the criterion in Eq. (5.17) requires the solution of the equation

$$Bv = \lambda W v. \tag{5.18}$$

The generalized eigenvalue problem is an extensively researched area in the linear algebra. Several numerical libraries are available which provide iterative algorithms solving different forms of this problem. In this work, the *LAPACK* library [Anderson et al. 1999] has been used.

In the following, important remarks on LDA are summarized with regard to its shortcomings and applicability:

- A restriction of the LDA approach is that the data distributions of each class have to share a common covariance matrix. The LDA transformation has been extended for heteroscedastic data in [Kumar and Andreou 1998].

- In [Pitz 2005], LDA is placed in a maximum likelihood framework. It can be shown that finding an optimal linear transformation of feature vectors by optimizing the maximum likelihood criterion leads to the generalized eigenvalue problem as well. However, note that a few assumptions have to be made: Gauss distribution of the observation vectors within a class and maximum approximation while processing all possible HMM state sequences (i.e. Viterbi alignment).

- Mostly, HMM states are chosen as classes to be separated by the LDA but many other classes are suitable. In fact, in the RWTH recognition system, the generalized HMM states (i.e. triphone state tying classes) have emerged as best choice in terms of word error rate.

- The generalized eigenvalue problem is invariant with respect to invertible transformations of the feature vectors. Nevertheless, scaling of acoustic features may become necessary if the numerical ranges are rather different. The same order of magnitude of the feature vector coefficients ensures numerically stable estimates of the scatter matrices.

- The generalized eigenvalue problem does not define the length of the eigenvectors. In fact, the length of the eigenvectors determines the projected diagonal within-class covariance matrix. Although, from a theoretical point of view, the length of the eigenvalues does not have any effect on the decision of the Bayes' decision rule, from a practical point of view, it is beneficial to ensure that

$$V_D^T W V_D = I. \tag{5.19}$$

It can be implemented by scaling the eigenvectors accordingly. The automatic application of such a scaling of the eigenvectors ensures that the order of magnitude of the acoustic scores is invariant with respect to any changes in the feature space as long as the output dimension of LDA is kept constant. Consequently, settings of the training and recognition systems given in terms of acoustic scores or variances do not need to be readjusted for every new acoustic feature setup. Such parameters are for instance pruning thresholds.

5.2 Feature Combination Algorithm

In the following steps, an application of LDA is described which combines acoustic features of extracted by different extraction algorithms in a straightforward way:

1. Feature vectors $x_{f_i}^{[t]}$ extracted by different algorithms f_i are concatenated for all time frames t.

2. The $2L + 1$ successive concatenated vectors are concatenated again for all time frames t which makes up the large input vector of LDA. With $L = 5$ and with $F = 3$ different features, the size of the LDA input vector $\mathcal{D}$ grows up to $\mathcal{D} \approx 400$ components.

3. The combined feature vector $y^{[t]}$ is created by projecting the large input vector on a smaller ($D \approx 40$ dimensional) subspace:

$$y^{[t]} = \left[\ V_D^T\ \right] \left[\begin{array}{c} \left[\begin{array}{c} x_{f_1}^{[t-L]} \\ \cdots \\ x_{f_F}^{[t-L]} \end{array}\right] \\ \cdots \\ \left[\begin{array}{c} x_{f_1}^{[t]} \\ \cdots \\ x_{f_F}^{[t]} \end{array}\right] \\ \cdots \\ \left[\begin{array}{c} x_{f_1}^{[t+L]} \\ \cdots \\ x_{f_F}^{[t+L]} \end{array}\right] \end{array}\right] \tag{5.20}$$

where the matrix V_D is determined by LDA such that it conveys the most relevant classification information to $y^{[t]}$.

The estimation of the LDA matrix V_D is performed in the standard way:

1. Estimation of the between- and within-class scatter matrices, B and W respectively. The LDA classes are chosen to be the generalized triphone states.

2. The generalized eigenproblem

$$Bx_j = \lambda W x_j \tag{5.21}$$

is solved by applying the corresponding functionality of the *LAPACK* library [Anderson et al. 1999].

3. Those eigenvectors x_j are collected in the columns of the LDA matrix V_D which belong to the first D largest eigenvalues. D is optimized empirically for each corpus.

In this way, not only the combination of successive time frames is carried out but an optimal combination of the underlying feature streams f_i is also realized.

5.3 Experimental Setup

For the different corpora, two parameters of the LDA based feature combination have been optimized: the number of concatenated successive feature vectors (L) and the dimension of the projected feature space, cf. Eq. (5.20). Both parameters have been optimized empirically using the MFCC feature. In Chapter A, Table A.1 summarizes the optimal settings. On average, an LDA window of ≈ 11 ($L = 5$) and a dimension after LDA of ≈ 40 seems to provide the best recognition results. The effects which influence the optimum of the LDA window length have not been found yet. In Section 5.5, different explanations are provided why increasing context does not lead to monotonously decreasing WER. The optimal number of output coefficients is influenced by the amount of training data. Since the LDA output coefficients are used as acoustic features in the complete training and recognition procedure, the number of them determines the dimension of the feature space and consequently the number of the emission model coefficients. The available training set has to be sufficiently large to train the parameters of the emission models reliably.

For a given corpus, the size of the projected feature vectors has been kept constant throughout different experiments to ensure comparable numbers of parameters and therefore comparability of recognition results. Nevertheless, the LDA input dimension increases with the number of feature sets combined, therefore implying a slight increase of parameters for the LDA transformation matrix. Finally, note that LDA has been applied in baseline experiments using a single feature in the same way as in feature combination experiments.

5.4 Recognition Results

In this section, experiments are presented using the LDA based combination of the features introduced in Chapter 4. The goal of these experiments is to test the ability of LDA to combine increasing number of different acoustic features. The results have turned out to be on the one hand satisfactory, i.e. either improvements or no significant changes in WER have been obtained. On the other hand, there have been features whose combination has yielded degradations in WER.

Table 5.1 shows improving recognition results obtained by the combination of the MFCC, vocal tract length normalized MFCC (VTLN), voicing (V), and spectrum derivative (SD) features. On the small-vocabulary *SieTill* task, an improvement in WER of 11% is obtained relative to the MFCCs when adding the voicing feature. Extending the set of features by the spectrum derivative feature improves the WER further by relative 6%. Nevertheless, this improvement has turned out to be significantly less than the improvements obtained by the direct combination of the MFCC and the spectrum derivative features. This observation has been confirmed on the large-vocabulary corpora as well. Hence, the relation between the voicing and the spectrum derivative features should be investigated further in the scope of feature combination.
On the large-vocabulary tasks, the best performing state-of-the-art system has been chosen as baseline. On the *VM II* and *EPPS* corpora, the combination of the VTLN and voicing features extended by the spectrum derivative feature results in a monotonous improvement of the WER. Finally, the MFCC feature has been added to test the robustness of LDA against increasing number of features and against increasing input size. No significant change in WER has been expected since the MFCC feature is strongly related to the VTLN one. As shown in Table 5.1, the additional MFCC feature has yielded neither a significant improvement nor a degradation in WER.

Table 5.1: Consistent improvements in word error rates obtained by the LDA based combination of increasing number of acoustic features.

corpus	acoustic feature	error rates [%] del	ins	WER
SieTill	MFCC	0.3	0.5	1.8
	+V	0.3	0.4	1.6
	+SD	0.2	0.3	1.5
VM II	VTLN	3.8	2.9	19.1
	+V	4.1	2.7	18.7
	+SD	3.9	2.9	18.4
	+MFCC	3.6	2.8	18.3
EPPS	VTLN	4.3/3.7	1.3/1.5	14.2/14.1
	+V	4.0/3.3	1.5/1.6	13.8/14.0
	+SD	3.6/3.1	1.6/1.8	13.7/14.0
	+MFCC	3.7/3.3	1.6/1.9	13.8/14.1

Besides consistent monotonous changes in WER, an unexpected degradation has been obtained when combining the MFCC, MF-PLP, and PLP features. Table 5.2 summarizes the baseline recognition results of the individual features and the results obtained by the LDA based combination denoted by Σ_{LDA}. On both corpora, strong degradations in WER have been obtained when combining the MFCC feature with weaker performing features. This observation does not comply with the results shown in Table 5.1. There, the combination of the VTLN, V, and SD feature set with the much weaker performing MFCC feature did not cause any significant degradation in WER. A possible explanation can be found if the correlation between the combined features is considered in detail.

In Section 5.5, the numerical aspects of the above experiments are analyzed along with the discussion of aforementioned open questions on the application of LDA to feature extraction.

Table 5.2: Degradation in word error rate obtained by the LDA based combination of baseline features MFCC, MF-PLP, and PLP.

corpus	acoustic feature	error rates [%] del	ins	WER
VM II	MFCC	4.5	2.9	21.0
	MF-PLP	5.2	2.3	21.0
	PLP	5.9	2.3	21.4
	Σ_{LDA}	4.7	3.3	21.6
EPPS	MFCC	4.3/3.8	1.4/1.7	14.7/15.3
	MF-PLP	4.2/3.7	1.5/1.7	14.8/15.3
	PLP	4.3/3.5	1.6/1.8	15.4/15.8
	Σ_{LDA}	4.8/4.1	1.4/1.6	15.8/16.2

5.5 Analysis of Results

Robustness of the application of LDA to acoustic feature extraction has already been addressed in [Welling 1999] and [Katz et al. 2002]. The authors have mainly addressed the problem of increasing input vector size caused by either increasing number of feature coefficients or by increasing LDA window length. In the following, conclusions of the two publications are summarized:

- In [Welling 1999], experiments have been presented with increasing LDA window length i.e. with increasing number of successive concatenated feature vectors. Instead of monotonous improvement in WER, a clear optimum has been found at ≈ 11 concatenated feature vectors. The author has proposed two explanations for the degradation in WER when increasing the LDA window length over the optimum:

  1. LDA is based on a maximum likelihood estimation using the generalized triphone states as classes to be optimally separated. Hence, the estimation of the LDA projection matrix is not bound directly to the word error rate.
  2. LDA operates on estimated values of the within- and between-class scatter matrices. The larger the context of the LDA input, the more scatter elements have to be estimated using a constant amount of training feature vectors. Unreliable estimates of the scatter matrices can have large impact on the quality of the solution of the eigenvalue problem.

- In [Katz et al. 2002], more specific experiments have been conducted to investigate the robustness of LDA applied to automatic speech recognition.

1. In the first kind of experiments, artificial data has been used to test the robustness of LDA against increasing feature dimension. The feature vectors have been augmented with increasing number of white noise components. In this simulation, the scatter matrices have been estimated on 100 training feature vectors per class. The classification error rate has doubled when augmenting the two-dimensional feature vectors with 200 for the classification irrelevant white noise components.
2. In the second kind of experiments, a real speech recognition system has been used to evaluate the robustness of LDA. The goal of the experiments has been to investigate the necessary amount of training data for increasing LDA window length. Recognition tests have shown that for increasing LDA window length, an increasing amount of training data is required to retain the best recognition result.

In the following, a discussion of the above observations is presented. Furthermore, a new aspect of robustness is introduced which has emerged from the combination of acoustic features. In Section 5.5.1, the experiments with white noise components are repeated on a real corpus rather than on artificial data to carry out stress tests by increasing the feature dimension. In order to find explanations for the degradation in WER presented earlier in Table 5.2, the sensitivity of the eigenvalue and eigenvector estimates are examined by means of the *perturbation theory* in Section 5.5.2. In addition to the discussion of numerical aspects, an analysis of the LDA based feature combination has been carried out in terms of recognition errors. In Section 5.5.3, confusion matrices generated by different feature setups are compared to evaluate the effect of feature combination on typical errors made by the RWTH small-vocabulary recognition system.

5.5.1 Combination with White Noise Components

LDA provides a simple way to combine acoustic feature vectors produced by different algorithms and in different time frames. Nevertheless, LDA yields unexpected results when increasing the window length or when combining certain features. In this section, the robustness of LDA is investigated against increasing input dimension. The analysis can be divided into two separate problems:

1. The problem of data sparseness caused by the increasing feature dimension has been addressed in [Katz et al. 2002]. Increasing the number of elements in the scatter matrices increases the amount of events that can occur form a statistical point of view. The authors recommend a minimum of 2000 training feature vectors per LDA class when using 160 dimensional LDA input vectors for a 120-class classification problem.

2. The increasing input dimension can lead to instabilities when solving the *generalized eigenvalue problem*. Depending on the implementation of the algorithms, large matrices may not be handled as precisely as smaller ones.

The goal in this section is to focus solely on the second aspect. If the degradations in WER are assumed to be caused by numerical instabilities during solving the generalized eigenvalue problem then additional artificial white noise components should be able to induce these problems respectively cause increasing WER. The experiments have been performed on large-vocabulary tasks using the MFCCs as baseline feature. The LDA window length has been kept constant throughout the experiments, 11 for the *VM II* and 9 for the *EPPS* corpus. In order to increase the input dimension of LDA and to stress the generalized eigenvalue problem, the baseline 16 MFCC components have been augmented with white noise components simulating additional features. In this way, an arbitrary amount of feature vector coefficient can be produced per time frame without changing the classification information content of the feature vectors. Hence, if the underlying eigenproblem is solved correctly, no change in the WER should be observed.

The eigenvalue problem of a singular scatter matrix pair is much more complicated than for a regular pair. Therefore, the white noise components are required to be independent of each other within one time frame and between time frames. Standard random number generators do not comply with this requirement. E.g. for a corpus of ≈ 60 hours of training data, using a 100 dimensional white noise extension for every MFCC feature vector requires a random number generator with a periodicity greater than 3.6×10^9. In the following experiments, a random number generator was used optimized for floating point precision implemented in [Press et al. 1992]. The *ran2* function ensures a minimum period of 2×10^{18} which abundantly fulfills the requirements of the processed corpora.

Recognition results are summarized in Table 5.3. The 16 MFCC components have been extended with up to 90 white noise coefficients per time frame. The resulting concatenated LDA input vectors have grown up to 954 components per time frame. For both corpora, the white noise components have not caused any degradation in WER. Unlike reported

Table 5.3: Word error rate obtained by the LDA based combination of MFCC features and increasing number of randomly generated coefficients.

# rnd cmp	# LDA input	error rates [%] del	ins	WER
VM II corpus				
0	176	4.5	2.9	21.0
15 ∗ 11	341	4.5	3.0	20.9
30 ∗ 11	506	4.5	3.0	20.9
38 ∗ 11	594	4.8	3.0	21.0
EPPS corpus				
0	176	4.3/3.8	1.4/1.7	14.7/15.3
30 ∗ 9	414	4.3/3.8	1.3/1.7	14.6/15.2
60 ∗ 9	684	4.2/3.9	1.4/1.7	14.6/15.2
90 ∗ 9	954	4.3/3.9	1.4/1.7	14.8/15.4

in [Katz et al. 2002], the experiments on real corpora have not led to changes in WER.

One important difference in the experimental setups has to be mentioned for sake of fair comparison. There is a large difference between the number of observations per LDA class. Using ≈ 4000 generalized triphone states as LDA classes, in the RWTH setup one class has had ≈ 4500 observations compared to 100 presented in [Katz et al. 2002]. Thus, at this stage of the analysis, the following conclusion can be drawn: the increasing feature vector size up to 106 components per time frame does not induce numerical instabilities in LDA when solving the generalized eigenvalue problem.

5.5.2 Sensitivity of Eigenvalues and Vectors

The application of LDA to feature combination resulted in as well monotonous improvements as degradations in WER when combining increasing number of features. In Section 5.5.1, the numerical instability of the eigenproblem caused by increasing input dimension has been ruled out as possible reason for the performance loss. In this section, another numerical aspect of the generalized eigenvalue problem is analyzed. By means of the perturbation theory, the reliability of the eigenvalue and eigenvector estimates is investigated while calculating the LDA projector matrix. The theory of the sensitivity or with other word reliability of eigenvalues and eigenvectors estimates is recalled in Section 5.5.2.1. In recognition experiments presented in Section 5.5.2.2, a potential relationship is verified between the sensibility of eigenvalue and eigenvector estimates and the word error rate.

5.5.2.1 Definition of Sensitivity of Eigenvalues and Vectors

One of the most important sources of numerical instability is the correlation between feature vector components. The stronger the correlation between the different features, the more singular the scatter matrices become. However, on real-life speech corpora, the correlation between feature components cannot be determined clearly. The transition between correlated and uncorrelated features is continuous, an obvious border cannot be found easily.

Most of the algorithms optimized for the symmetric generalized eigenproblem assumes a positive definite within-class scatter matrix. In order to analyze numerical instabilities caused by correlation of features, the within-class scatter matrix is not assumed to be positive definite. Allowing singularity in the eigenproblem requires more complex algorithms than the definite case. Generally, the following differences can be found between singular (indefinite) and positive definite (shortly definite) matrix pencils $(A - \lambda B)$:

- An indefinite matrix pencil $(A - \lambda B)$ of the symmetric matrix pair (A, B) may lead to complex eigenvalues, in contrast to a definite one yielding only real eigenvalues.
- An indefinite matrix pencil may not have a complete set of eigenvectors.
- If both A and B are singular, or close to singular, then any complex number λ is an eigenvalue.

In all experiments, the linear algebra software library *LAPACK* [Anderson et al. 1999] has been applied to the generalized eigenvalue problem in LDA. Although, several algorithms are available optimized for generalized symmetric definite eigenproblems, due to allowing indefinite matrix pencils, the *dggevx* algorithm has been applied which has been developed for generalized nonsymmetric eigenproblems. The algorithm can be summarized as follows:

$$\begin{aligned} Bx_i &= \lambda_i W x_i \\ y_i^H B &= \lambda_i y_i^H W \end{aligned} \tag{5.22}$$

where B and W denote the within- and between-class scatter matrices. The left and right eigenvectors are denoted by y_i and x_i, respectively, belonging to the eigenvalue λ_i. After reducing (B, W) to the generalized upper Hessenberg form, the generalized Schur decomposition is applied to calculate the upper triangular matrix pair (S, T). The two orthogonal transformations are unified in Q and Z

$$\begin{aligned} S &= Q^T B Z \\ T &= Q^T W Z. \end{aligned} \tag{5.23}$$

The left and the right eigenvectors are computed based on the eigenvectors of the upper triangular matrix pair (S, T). The eigenvalues λ_i are calculated from the diagonal elements of (S, T) as follows,

$$\lambda_i = \frac{S_{ii}}{T_{ii}} = \frac{\alpha_i}{\beta_i}. \tag{5.24}$$

Before calculating the eigenvalues, extreme cases can be found by checking (α_i, β_i) for values close to zero. Note that although singularities can be found in this way, the reliability of the rest of the eigenvalue and vector estimates remains heavily effected by the singularity(ies), as shown experimentally in Section 5.5.2.2.

After defining the eigenproblem included in LDA, an asymptotic error bounds can be introduced for the eigenvalue and eigenvector estimates. Assume that (α, β) make up a real eigenvalue $\lambda = \frac{\alpha}{\beta}$ of the matrix pair (B, W) with left and right eigenvectors y and x, respectively, satisfying Eq. (5.22). The estimated eigenvalues (α', β') can be seen as real eigenvalues of the perturbed matrix pair $(B + E, W + F)$ where $||(E, F)|| = \epsilon||(B, W)||_1$ and ϵ is the 64bit machine precision. In the perturbation theory, for generalized eigenvalues the *chordal distance* between (α, β) and (α', β') is considered. The chordal distance is defined in the following way:

$$\mathcal{X}((\alpha, \beta), (\alpha', \beta') = \frac{|\alpha\beta' - \alpha'\beta|}{\sqrt{|\alpha|^2 + |\beta|^2}\sqrt{|\alpha'|^2 + |\beta'|^2}}. \tag{5.25}$$

Note that this definition allows the interpretation of the case of a singular W. Instead of speaking of small perturbation to W causing small perturbation to an infinite value, the perturbations to α and β is addressed which have values, in case of a singularity, close to zero. Then, an asymptotic upper bound for the error between the real and the estimated eigenvalues is given in the form of

$$\mathcal{X}((\alpha, \beta), (\alpha', \beta') \leq \frac{\epsilon||(B, W)||_1}{S(\lambda)}, \tag{5.26}$$

where $S(\lambda)$ is called the reciprocal condition number of the eigenvalue λ. Small values of $S(\lambda)$ indicate ill-conditioned eigenvalues, since a small perturbation of the matrix pair (B, W) results in a large difference between the estimated and the real eigenvalues. Note that beside the reciprocal condition number, the 1-norm of the matrix pair (B, W) influences also the error bound. Increasing dimension of (B, W) may lightly lead to increasing 1-norm and consequently to increasing error bound.

In order to define the asymptotic error bound and the reciprocal condition number of a left and right generalized eigenvector pair, the properties of the *deflating subspaces* of the matrix pair (B, W) are reviewed in the following. Deflating subspaces are the generalization of the invariant subspaces of a single matrix. Let us short recall the definition of an invariant subspace $\mathcal{X}$ of a transformation A:

$$A\mathcal{X} \subset \mathcal{X}. \tag{5.27}$$

In case of the generalized eigenproblem, a pair of right and left deflating subspaces $\mathcal{X}$ and $\mathcal{Y}$ of the pencil $A - \lambda B$ are defined as:

$$A\mathcal{X} + B\mathcal{X} \subset \mathcal{Y}. \tag{5.28}$$

The left and right eigenvectors as well as the first successive k columns $\{[q_1...q_k], [z_1...z_k]\}$ of (Q, Z) form left and right deflating subspaces of (B, W). Therefore, conditioning of individual eigenvectors can be regarded as special case of conditioning for left and right deflating subspaces. Let us define general one-dimensional left $\mathcal{L} = span\{q_1\}$ and right $\mathcal{R} = span\{z_1\}$ deflating subspaces of (B, W) belonging to the eigenvalue $\lambda = \frac{\alpha}{\beta}$ and to the left and right eigenvectors (y, x), respectively:

$$\begin{aligned} [q_1\ Q_2]^H B[z_1\ Z_2] &= S = \begin{bmatrix} \alpha & S_{12} \\ 0 & S_{22} \end{bmatrix} \\ [q_1\ Q_2]^H W[z_1\ Z_2] &= T = \begin{bmatrix} \beta & T_{12} \\ 0 & T_{22} \end{bmatrix}. \end{aligned} \tag{5.29}$$

Furthermore, let the estimated one-dimensional $\mathcal{L}' = span\{q_1'\}$ and $\mathcal{R}' = span\{z_1'\}$ left and right deflating subspaces be the real deflating subspaces of the perturbed matrix pair $(B+E, W+F)$ where $||(E, F)|| = \epsilon||(B, W)||_1$ and ϵ is the 64bit machine precision. In the perturbation theory, a frequently used distance between two one-dimensional subspaces spanned by u and v is defined as

$$\theta(u, v) = \arccos \frac{|u^T v|}{||u||_2||v||_2}. \tag{5.30}$$

Then, the following first order bounds define the asymptotic error bounds for the eigenvector pair (x, y)

$$\begin{aligned} \theta(\mathcal{L}, \mathcal{L}') &\leq \tfrac{\epsilon||(B,W)||_1}{Dif_l(\lambda,x,y)} \\ \theta(\mathcal{R}, \mathcal{R}') &\leq \tfrac{\epsilon||(B,W)||_1}{Dif_l(\lambda,x,y)}, \end{aligned} \tag{5.31}$$

along with the reciprocal condition number $Dif_l(\lambda, x, y)$. Small values of $Dif_l(\lambda, x, y)$ indicate ill-conditioned eigenvectors, since a small perturbation of the matrix pair (B, W) results in a large difference in the angles between the original and the perturbed deflated subspaces. Note that similar to the case of the eigenvalues, the 1-norm of the matrix pair (B, W) influences the eigenvector error bound as well. Furthermore, since the asymptotic error bound is an upper bound for the angle between two vectors, values larger than π have no significance. More details on the motivation and estimation of the condition numbers can be found in [Golub and Loan 1996].

5.5.2.2 Experimental Results

After introducing the problem of singularities during the estimation of the LDA matrix, recognition results are presented which have been obtained on the *EPPS* corpus. The purpose of the experiments was to analyze the relationship between the word error rate and the asymptotic error bounds respectively the reciprocal condition numbers. For each experiment, the *arithmetic average* of the error bounds and of the reciprocal condition numbers is presented calculated over all eigenvalues respectively eigenvectors. Table 5.4 summarizes results obtained by combining different sets of features. The first line presents the baseline results using the MFCC feature. The next experiment has been performed to stress the LDA matrix estimation by an artificially introduced singularity. In every time frame, the first MFCC coefficient has been repeated simulating a highly correlated additional one-dimensional feature. Although the information content of the acoustic features has not changed, the WER decreased by 6% relative to the baseline results using the same features. Simple methods, like explicitly excluding eigenvectors from the projection matrix which belong to $(\alpha, \beta) < \mu$ has not improved the results. Several values of μ have been tested, the best recognition result has been obtained by setting μ to 3×10^{-8}. A possible explanation for the degradation in WER can be found when comparing the conditioning of this eigenvalue problem to the baseline experiment. The average condition numbers have dropped rather heavily, indicating weak estimates of the eigenvalues and eigenvectors. Furthermore, the low average condition number indicates that a strong singularity effects not only the conditioning of the singular eigenvalue but the rest of the eigenvalue estimates become badly conditioned as well. The third and fourth lines of the table show results of experiments using the combination of different features. As expected, the experiment combining MFCC, VTLN, V, and SD features has not resulted in significantly different in the condition numbers compared to the baseline experiment. Observe that the eigenvector condition number has decreased by an order of magnitude of 10^{-2} what should become a subject of future investigations. Finally, the average reciprocal condition numbers have been calculated for the combination of the MFCC, MF-PLP, and PLP features. Our goal was to find an explanation for the unexpected increase in WER compared to using the single MFCC feature. Although the degradation in WER is comparable with the experiment repeating the first MFCC coefficient, the condition numbers have not decreased as heavily as in the second line of the table. Further analysis of the problem is required to verify if the small reduction in the eigenvalue condition number and the increase in the asymptotic error bounds explain the degradation in WER.

Table 5.4: Average reciprocal condition numbers and asymptotic error bounds of eigenvalues and eigenvectors on the *EPPS* corpus obtained by LDA based feature combination tests yielding improvements and degradations in word error rate.

acoustic features	WER [%]	#LDA input	avr. recip. cond. num. eigenvalue	avr. recip. cond. num. eigenvector	avr. asym. error bound eigenvalue	avr. asym. error bound eigenvector
MFCC	14.7/15.3	144	4.0	2.7×10^{-2}	2.5×10^{-13}	3.7×10^{-8}
MFCC+Repeated-Coeff	15.6/16.2	153	1.7×10^{-8}	2.0×10^{-19}	8.3×10^{2}	$> \pi$
MFCC+VTLN+V+SD	13.8/14.1	306	1.0	5.2×10^{-4}	2.3×10^{-11}	1.0×10^{-4}
MFCC+MF-PLP+PLP	15.8/16.2	432	1.5×10^{-1}	6.8×10^{-4}	3.0×10^{-11}	2.5×10^{-3}

Table 5.5, summarizes results obtained by increasing the LDA window length i.e. the number of successive concatenated feature vectors. The purpose of the experiments was to verify if the increasing size of the scatter matrices leads to ill-conditioned eigenproblems. The analysis of the condition numbers has not confirmed our hypothesis. The condition numbers do not change significantly when increasing the LDA window length. Although the asymptotic error bounds do show an increasing tendency, their relation to the word error rate is not obvious. The definitions Eq. (5.26) and Eq. (5.31) indicate that the increasing error bounds must be caused by the increasing matrix norm $||(B, W)||_1$ in case of constant reciprocal condition numbers.

Table 5.5: Average reciprocal condition numbers and asymptotic error bounds of eigenvalues and eigenvectors on the *EPPS* corpus obtained by using increasing LDA window lengths.

window length	WER [%]	#LDA input	avr. recip. cond. num. eigenvalue	avr. recip. cond. num. eigenvector	avr. asym. error bound eigenvalue	avr. asym. error bound eigenvector
5	15.5/16.8	80	4.3	3.4×10^{-2}	1.4×10^{-13}	2.5×10^{-9}
7	15.0/15.3	112	4.2	3.2×10^{-2}	2.0×10^{-13}	9.9×10^{-9}
9	14.7/15.3	144	4.0	2.7×10^{-2}	2.5×10^{-13}	3.7×10^{-8}
11	15.0/15.5	176	4.0	2.1×10^{-2}	3.5×10^{-13}	1.2×10^{-7}
13	15.1/15.6	208	4.1	2.2×10^{-2}	3.9×10^{-13}	2.0×10^{-7}
17	15.4/15.8	272	3.8	2.1×10^{-2}	2.0×10^{-12}	1.2×10^{-5}

5.5.3 Comparison of Confusion Matrices

The fundamental goal of feature combination is to utilizing mutually complementary classification information of different acoustic features. In the following, the LDA based feature combination is studied in terms of changes in recognition error. In ideal case, adding a new acoustic feature should correct recognition errors, leave correctly recognized words unchanged, or may of course change misrecognized words. The case of changing a

correctly recognized word indicates that the feature combination algorithm does not make optimal use of the mutually complementary information coded in the underlying features.

Errors produced by speech recognition systems can be illustrated by using confusion matrices. An element c_{ij} of a confusion matrix C gives how many times the word i has been spoken and word j has been recognized. Naturally, the sum of the diagonal elements gives the number of correctly recognized words and off-diagonal elements represent recognition errors. Table 5.6 shows the difference of two confusion matrices. Both matrices have been obtained on the small-vocabulary digit string recognition *SieTill* task. In order to show the changes caused by using the additional voicing feature, the difference of the confusion matrices is presented rather than showing the single matrices separately. The confusion matrix obtained by using the LDA combined MFCC and voicing features has been *subtracted* from the one generated by using solely the MFCC feature. Consequently, in the diagonal of the difference confusion matrix, negative elements show improvements and positive ones degradations. Naturally, negative off-diagonal elements indicate degradations and positive ones improvements. Larger **improvements** and *degradations* are emphasized in Table 5.6. Although the overall number of errors has decreased,

Table 5.6: Difference confusion matrix generated by subtracting the confusion matrix obtained by using the LDA based combination of the MFCC and voicing features from the one obtained by using solely the MFCC feature.

	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	'0'	'zwo'
'1'	**-10**	3	**5**	0	2	2	0	0	3	0	1
'2'	3	**-17**	*-16*	0	0	0	0	0	0	0	-2
'3'	1	**9**	*11*	0	-1	1	0	0	-1	1	0
'4'	0	-1	-1	-3	0	0	3	-1	-1	0	0
'5'	**5**	0	1	0	**-7**	1	0	0	0	0	-3
'6'	0	0	-1	0	3	**-6**	0	2	0	-1	0
'7'	1	0	-1	0	2	1	**-5**	1	-1	1	0
'8'	0	0	-1	1	0	0	2	0	0	0	0
'9'	1	0	3	0	-3	0	0	1	**-11**	*-6*	1
'0'	0	3	-1	-1	0	0	1	0	**11**	-1	-3
'zwo'	0	3	1	0	2	1	-1	-1	-1	2	*6*

larger degradations can be found in the difference matrix. The words "3" and "zwo" have been misrecognized more often after adding the voicing feature. Furthermore, the amount of confusions between the words "2" and "3" as well as "9" and "0" has increased considerably. Hence, the integration of the voicing feature is not optimal. Although the amount of improvements is greater than the amount of degradations, an optimal integration of the voicing feature should not have resulted in degradations in the recognition accuracy of any word.

As mentioned in Section 5.1, the training of the LDA matrix is not directly bound to recognition errors. Further improvements may be obtained if discriminative methods were applied to the acoustic feature combination. One of the straightforward extensions

of the LDA based feature combination is to estimate the LDA matrix by optimizing a discriminative training criterion.

5.5.4 Summary

The conclusions of the analysis of LDA presented in this section can be summarized as follows:

- Experiments with additive white noise components indicate that the LDA based feature combination and the underlying eigenproblem is robust against scatter matrices of up to ≈ 1000 dimension.
- When combining different acoustic features, stronger correlation between the feature vector components leads to ill-conditioned eigenvalue problems and consequently degradation in WER. Algorithms developed for general singular pencils [Demmel and Kagström 1993a][Demmel and Kagström 1993b] or for symmetric ones [Parlett and Chen 1990] should be used to improve the eigenvalue and eigenvector estimates.
- One of the most possible explanation for the degradation in WER when combining the MFCC, MF-PLP, and PLP features and when increasing the LDA window length remains the insufficient amount of data to train the increasing scatter matrices.
- The combination of the MFCC and voicing features has yielded degradation in the recognition accuracy of certain words on the *SieTill* corpus. An improvement of the LDA based combination of these features may be expected if a discriminative criterion were used rather than the maximum likelihood one to estimate the LDA matrix.
- LDA can be formulated in a probabilistic framework as well. Consequently, making use of a-priori probabilities may improve the stability of the method.

Chapter 6

DMC Based Feature Combination

The Discriminative Model Combination (DMC) has been first proposed in [Beyerlein 1997]. This method provides a flexible framework for combining numerous knowledge sources of different kinds. In [Beyerlein 1998], the combination of several acoustic models with different allophone context lengths and several language models has been presented yielding improvement in WER of 5% on a large-vocabulary tasks compared to the best pairwise combinations. One of the first attempts at applying DMC to feature combination was presented briefly in [Häb-Umbach and Loog 1999]. The combination of three state-of-the-art acoustic features (MFCC, MF-PLP, and LPC-smoothed MFCC) has been tested on the Hub-4 corpus. An improvement in WER of 3% has been obtained relative to using the MFCC feature alone. In the following, a description of the application of DMC to feature combination is presented based on a splitting of the acoustic model into separate scaleable knowledge sources. The implementation of the RWTH DMC toolkit has been accomplished in cooperation with Georg Heigold.

The remainder of this chapter is organized as follows: The DMC based feature combination algorithm is introduced in Section 6.1. The training of knowledge sources and log-linear model weights is discussed in Section 6.2. Integration of the log-linear model into the recognition process is described in Section 6.3. Finally, recognition tests are presented in Section 6.4 which aimed at the combination of several auditory and articulatory motivated acoustic features.

6.1 Feature Combination Algorithm

In this approach, different acoustic features are combined indirectly via the log-linear combination of acoustic models $P_{f_i}(X_{f_i}|W)$ where W denotes a sequence of words and X_{f_i} denotes a sequence of feature vectors extracted by the algorithm f_i. Whereas in the LDA based method, the feature combination matrix is estimated in one of the first steps of the acoustic training, the DMC based method combines the probabilities produced by complete acoustic models trained on the different features. The combination of acoustic probabilities is performed in a log-linear manner where the model weights are trained by using a discriminative training criterion, e.g. Minimum Word Error.

The basic idea of DMC is to modify the modeling of the a-posteriori probability $P(W|X)$ in the Bayes' decision rule:

$$W_{opt} = \underset{W}{\operatorname{argmax}}\, P(W|X). \tag{6.1}$$

In the standard case, a-posteriori probability is decomposed into language model probability $P(W)$ and acoustic model probability $P(X|W)$:

$$P(W|X) = \frac{P(W)P(X|W)}{\sum_{W'} P(W')P(X|W')}. \tag{6.2}$$

In the case of Discriminative Model Combination, the a-posteriori probability is modeled by a log-linear distribution:

$$P_\Lambda(W|X) = \frac{e^{\sum_i \lambda_i g_i(W,X)}}{\sum_{W'} e^{\sum_i \lambda_i g_i(W',X)}} \tag{6.3}$$

where g_i is a so called *feature function* which is an arbitrary function of the word sequence W and the feature vector sequence X, and λ_i is the corresponding log-linear weight. When applying the log-linear modeling approach to speech recognition, the basic feature function types are negative logarithm of probabilities:

- language model: $g_{lm}(W,X) := -\log P(W)$,
- acoustic model: $g_{am}(W,X) := -\log P(X|W)$.

In order to combine different acoustic features, X is redefined to be a sequence of tuples of vectors containing the time-synchronous acoustic feature vectors $x_{f_i}^{[t]}$ for time frames $t = 1, ..., T$:

$$X = \left\{ \begin{bmatrix} x_{f_1}^{[1]} \\ \cdots \\ x_{f_i}^{[1]} \\ \cdots \\ x_F^{[1]} \end{bmatrix} \quad \cdots \quad \begin{bmatrix} x_{f_1}^{[t]} \\ \cdots \\ x_{f_i}^{[t]} \\ \cdots \\ x_F^{[t]} \end{bmatrix} \quad \cdots \quad \begin{bmatrix} x_{f_1}^{[T]} \\ \cdots \\ x_{f_i}^{[T]} \\ \cdots \\ x_F^{[T]} \end{bmatrix} \right\} \tag{6.4}$$

Furthermore, separate acoustic feature functions $g_{am,f_i}(W,X)$ are introduced for each acoustic feature f_i. Theoretically, every feature function receives all the different acoustic feature vectors. Nevertheless, in the RWTH system, the acoustic feature functions make only use of the underlying acoustic feature f_i:

$$g_{am,f_i}(W,X) = g_{am,f_i}(W,X_{f_i}) = -\log P_{f_i}(X_{f_i}|W). \tag{6.5}$$

After introducing the basic feature function $g_{am,f_i}(W,X)$ of the acoustic feature f_i, it is split into smaller partially acoustic feature independent and partially acoustic feature dependent feature functions. The separation of the acoustic model into independently

scalable feature functions requires however the use of maximum approximation at two stages of the recognition process: the summation over all possible pronunciation variant sequences Φ and the summation over all possible state sequences S:

$$\begin{aligned} -\log P_{f_i}(X_{f_i}|W) &= -\log \sum_{\{\Phi\},\{S\}} P_{f_i}(X_{f_i}, S, \Phi|W) && (6.6)\\ &= -\log \sum_{\{\Phi\},\{S\}} \{P_{f_i}(X_{f_i}|S)P(S|\Phi)P(\Phi|W)\} && (6.7)\\ &\approx -\log \max_{\Phi,S} \{P_{f_i}(X_{f_i}|S)P(S|\Phi)P(\Phi|W)\} && (6.8)\\ &\approx \min_{\Phi,S} \{-\log P_{f_i}(X_{f_i}|S) - \log P(S|\Phi) - \log P(\Phi|W)\}\,. && (6.9) \end{aligned}$$

Since maximum operation and logarithm are commutative, the acoustic feature function can be replaced by the acoustic feature independent feature functions

- pronunciation model: $g_{pr}(\Phi, W) := -\log P(\Phi|W)$,
- transition model: $g_{tr}(S, \Phi) := -\log P(S|\Phi)$

and by the acoustic feature dependent feature function

- emission model: $g_{em,f_i}(S, X) := -\log P_{f_i}(X_{f_i}|S)]$.

Note that state-of-the-art search methods are already using these maximum approximations. Hence, this derivation should only underline that the splitting of the acoustic model feature function requires the use of maximum approximation in the recognition process.

The transition model $P(S|\Phi)$ can be split further into two types of transition: transition within word boundaries $P_{tr_w}(S|\Phi)$ and across word boundaries $P_{wrd}(S|\Phi)$. The first model defines the probability of transitions within a word, whereas the second model sets the probability of transitions across word boundaries.

Consequently, the Bayes' decision rule for log-linear feature combination using a single language model, pronunciation model, word penalty model, within word transition model, and for each acoustic feature a separate emission model can be written as:

$$\begin{aligned} W_{opt} = \operatorname*{argmax}_W \max_{\Phi,S} \{ & \\ & P(W)^{\lambda_{lm}} \cdot P(\Phi|W)^{\lambda_{pr}} \cdot \\ & P_{wrd}(S|\Phi)^{\lambda_{wrd}} \cdot P_{tr_w}(S|\Phi)^{\lambda_{tr_w}} \cdot \\ & \prod_i P_{f_i}(X_{f_i}|S)^{\lambda_{f_i}}\}. \end{aligned} \tag{6.10}$$

6.2 Training of DMC Based Feature Combination

The training of a DMC system consists of two major steps: independent training of the parameters of each feature function $g_i(W, X)$ and discriminative training of the log-linear

model weights λ_i. In this thesis, negative logarithms of probabilities have been used as feature functions, as described in the previous section. In order to train the parameters of the language, pronunciation, transition, and emission model distributions, standard maximum likelihood training has been performed. (Note that other training criteria could have also been used without any change in the DMC framework.) As introduced formerly, the DMC based combination of acoustic features does not require any special training of the emission feature functions either. For each feature extraction algorithm, a standard acoustic model is estimated using maximum likelihood training. Special parameters of the acoustic models can be optimized for example empirically.

The model weights are trained by using a discriminative criterion. Note that the training procedure is not specialized to acoustic feature combination. With such a training method, model weights of an arbitrary set of knowledge sources can be estimated.

The remainder of this section is organized as follows: The RWTH implementation of the maximum likelihood emission model training is described in detail in Section 6.2.1. The training of the model weights based on the *Minimum Word Error* criterion is presented Section 6.2.2.

6.2.1 Training of Emission Feature Function

The emission feature function $g_{em,f}(S, X) = -\log P_f(X_f|S)$ of an acoustic feature f is trained by using the maximum likelihood criterion. The training procedure does not only consists of the estimation of a Gaussian Mixture Model (GMM) set, but it also contains the estimation of a triphone state tying tree and the estimation of a LDA projection matrix. This section reviews the sequence of training actions rather than explaining the underlying algorithms. A detailed descriptions on the estimation of a LDA matrix can be found in Chapter 5. In [Beulen 1999], the RWTH implementation of triphone state tying is presented along with the optimization of the algorithm. The GMM set is trained by using the standard *Expectation Maximization* algorithm. The RWTH emission model training applies maximum approximation at two stages: summation over all possible alignments is replaced by considering only to best one, and the summation within a GMM is also replaced by considering only the best Gaussian density.

The remainder of this section is organized as follows: In Section 6.2.1.1, the initialization procedure of the emission model training is presented followed by the description of the training algorithm in Section 6.2.1.2. Finally, advantages of iterating the complete training algorithm are studied in Section 6.2.1.3.

6.2.1.1 Initialization

The emission model training presented in the following section requires two inputs: Classification and Regression Tree $CART_0$ and initial alignment $\mathcal{A}_0$. In the following, an algorithm is described which prepares these two objects passing several times over a training corpus. The algorithm starts with creating the first alignment by using linear segmentation. This alignment is first improved by several iterations using a simple monophone emission model consisting of single Gaussian densities. After estimating the

triphone state tying tree $CART_0$, the initial alignment is improved further by using a triphone based emission model. The algorithm is implemented in the RWTH training system as follows:

1. Generation of frame energy measures.
2. Generation of the start alignment $\mathcal{A}_{lin-seg}$ by using linear segmentation based on frame energy measures.
3. Caching of MFCC coefficients augmented with the first derivative and the second derivative of the first MFCC coefficient.
4. Improvement of the start alignment $\mathcal{A}_{lin-seg}$ by using monophone based emission model consisting of single Gauss densities $\mathcal{G}_{mono} \rightarrow \mathcal{A}_{mono}$:

   INPUT: $\mathcal{A}_{lin-seg}$
   OUTPUT: $\mathcal{A}_{mono}$
   ACTION:
   1. Estimate single densities $\mathcal{G}_{mono}$ based on $\mathcal{A}_{lin-seg}$.
   2. REPEAT 10×
      1. Determine new alignment based on $\mathcal{G}_{mono}$ and overwrite $\mathcal{A}_{mono}$.
         Remark: in each iteration, pruning needs adjustment to control runtime.
      2. Estimate single densities based on $\mathcal{A}_{mono}$ and overwrite $\mathcal{G}_{mono}$.

5. Estimation of triphone state tying tree $CART_0$.
6. Training of triphone single densities based start alignment $\mathcal{A}_0$:

   INPUT: $\mathcal{A}_{mono}$, $CART_0$
   OUTPUT: $\mathcal{A}_0$
   ACTION:
   1. Estimate single densities $\mathcal{G}_0$ based on $\mathcal{A}_{mono}$.
   2. REPEAT 10×
      1. Determine new alignment based on $\mathcal{G}_0$ and overwrite $\mathcal{A}_0$.
      2. Estimate single densities based on $\mathcal{A}_0$ and overwrite $\mathcal{G}_0$.

In summary, a short algorithm has been presented to generate initial values for an emission model training. In order to keep the initialization process simple, only the standard MFCC feature is used. Further simplification results from replacing the LDA by using the standard dynamic features. The initial CART tree $CART_0$ and initial alignment $\mathcal{A}_0$ can been used to start up the training of every emission model training independently of the acoustic feature.

6.2.1.2 Training Algorithm

The goal of the emission model training is to produce three objects which are required by different recognition or model combination systems: CART tree, LDA matrix, GMM set. In the following, the most recent RWTH system is presented which has been used to train the emission models tested in the scope of the this thesis. The algorithm requires two inputs which are provided by an initialization algorithm: initial triphone state tying tree $CART_0$ and the initial alignment $\mathcal{A}_0$. The algorithm is implemented as follows:

1. Caching acoustic single or concatenated acoustic feature vectors.
2. Estimation of the first LDA matrix LDA_1 base on $CART_0$ and $\mathcal{A}_0$.
3. Estimation of a new triphone state stying tree $CART$ based on $\mathcal{A}_0$ and on the output vectors of the projection with LDA_1.
4. Estimation of the second LDA matrix LDA_2 based on $\mathcal{A}_0$ and on the newly estimated $CART$.
5. Caching output vectors of the projection with LDA_2.
6. Training of the emission GMM set $\mathcal{G}_7$ based on $CART$ and on cached LDA_2 outputs:
   1. Training of single densities $\mathcal{G}_0$:

      INPUT: $\mathcal{A}_0$
      OUTPUT: $\mathcal{G}_0$
      ACTION:
      REPEAT 10×
      1. Estimate single densities based on $\mathcal{A}_0$ and overwrite $\mathcal{G}_0$.
      2. Determine new alignment based on $\mathcal{G}_0$ and overwrite $\mathcal{A}_0$.

   2. FOR $i = 1, \ldots, 7$:
      1. Spitting:

         INPUT: $\mathcal{G}_{i-i}$, $\mathcal{A}_{i-1}$
         OUTPUT: $\mathcal{G}_i$
         ACTION:
         1. Split each density in $\mathcal{G}_{i-i}$ with more than 20 observations $\rightarrow \mathcal{G}_i$.
            - Copy mean vector and perturbed both the original and the copy.
            - Note: both densities share the same unchanged weight.
         2. Reassign observations within the split GMMs in $\mathcal{G}_i$ based on $\mathcal{A}_{i-1}$.

      2. Reassignment:

         INPUT: $\mathcal{G}_i$, $\mathcal{A}_{i-1}$
         OUTPUT: $\mathcal{G}_i$ (overwritten)
         ACTION:
         Reassign observations within the GMMs in $\mathcal{G}_i$ based on $\mathcal{A}_{i-1}$.

3. Alignment:

   INPUT: $\mathcal{G}_i$
   OUTPUT: $\mathcal{A}_i$, $\mathcal{G}_i$ (overwritten)
   ACTION:
   1. Determine new alignment $\mathcal{A}_i$ based on $\mathcal{G}_i$.
   2. Reassign observations within the GMMs in $\mathcal{G}_i$ based on the new alignment $\mathcal{A}_i$.

4. Reassignment:

   INPUT: $\mathcal{G}_i$, $\mathcal{A}_i$
   OUTPUT: $\mathcal{G}_i$ (overwritten)
   ACTION:
   Reassign observations within the GMMs in $\mathcal{G}_i$ based on $\mathcal{A}_i$.

5. Alignment:

   INPUT: $\mathcal{G}_i$
   OUTPUT: $\mathcal{A}_i$ (overwritten), $\mathcal{G}_i$ (overwritten)
   ACTION:
   1. Determine new alignment $\mathcal{A}_i$ based on $\mathcal{G}_i$.
   2. Reassign observations within the GMMs in $\mathcal{G}_i$ based on the new alignment $\mathcal{A}_i$.

7. Output $CART$, LDA_2, and $\mathcal{G}_7$

In summary, a standard training procedure has been presented based on maximum approximation. The algorithm can be applied to any acoustic feature even to concatenated ones due to starting up directly with the estimation of LDA_1 matrix. The initial state tying $CART_0$ has not been estimated on the actual underlying acoustic features thus a new CART tree is estimated. Because of the interdependence between LDA and CART ,a second LDA matrix is estimated as well. The training of the GMM set is performed by alternating three main operations: splitting, alignment, and reassignment.

6.2.1.3 Iteration of Training Algorithm

The first steps of the emission training algorithm presented in the previous section rely on an initial alignment and CART tree. The initial objects are provided by a simplified training algorithm based on the standard dynamic MFCC features. Improvements in the emission model and consequently in WER can be obtained by iterating the complete training procedure. Every new iteration can reuse the $CART$ tree and the alignment $\mathcal{A}_7$ of the previous iteration as initial values. The iteration of the emission training algorithm can be defined as follows:

$\{CART_0, \mathcal{A}_0\}$ = initial-training();
$\{CART^{(0)}, \mathcal{A}_7^{(0)}, LDA_2^{(0)}, \mathcal{G}_7^{(0)}\}$ = emission-model-training($CART_0$, $\mathcal{A}_0$);
$WER_{dev}^{(0)}$ = recognition($CART^{(0)}$, $LDA_2^{(0)}$, $\mathcal{G}_7^{(0)}$);

$i = 0$;
DO
$i = i + 1$;
$\{CART^{(i)}, \mathcal{A}_7^{(i)}, LDA_2^{(i)}, \mathcal{G}_7^{(i)}\}$ = emission-model-training($CART^{(i-1)}$, $\mathcal{A}_7^{(i-1)}$);
$WER_{dev}^{(i)}$ = recognition($CART^{(i)}$, $LDA_2^{(i)}$, $\mathcal{G}_7^{(i)}$);
WHILE ($WER_{dev}^{(i)} < WER_{dev}^{(i-1)}$);

The iteration is carried on as long as the WER decreases on a development corpus. In this way, the estimation of LDA and CART can be improved by providing improved initial CART and alignment objects.

Experimental tests have shown that maximum two additional iterations of the complete training algorithm are necessary to obtain the best recognition results on the *VM II* corpus. On the *EPPS* corpus, the first pass results already in the best recognition results thus no further iterations are required.

6.2.2 Training of Model Weights

A discriminative training method to estimate log-linear model weights has been first presented in [Beyerlein 1997]. In the following, the RWTH implementation of this method is presented: In Section 6.2.2.1, the *Minimum Word Error* (MWE) criterion is introduced followed by the derivation of a gradient decent optimization algorithm. The appropriate representation of knowledge sources is described in Section 6.2.2.2. Finally, the initialization of the training process is studied in Section 6.2.2.3.

6.2.2.1 Training Algorithm

In the following, a discriminative training algorithm is presented which estimates weights in a log-linear model, see Eq. (6.3). The algorithm is based on the *Minimum Word Error* (MWE) criterion as described in [Povey and Woodland 2002, Macherey et al. 2005, Beyerlein 2000]. The MWE criterion is defined as follows:

$$F(\Lambda) = \sum_{r=1}^{R} \frac{\sum\limits_{V \in \mathcal{M}_r} [P_\Lambda(V|X_r)]^\eta \;\; \mathcal{L}(V, W_r)}{\sum\limits_{V' \in \mathcal{M}_r} [P_\Lambda(V'|X_r)]^\eta}. \tag{6.11}$$

This criterion minimizes the expected number of word errors over a development corpus w.r.t the model weights Λ. The development corpus is given by utterances r with $r = 1, \ldots, R$. W_r and X_r denote the sequence of words and the corresponding sequence

of tuples of synchronized acoustic feature vectors of the utterance r, respectively. The number of word errors (i.e. Levenshtein distance) between a recognized word sequence V and the uttered one W_r is denoted by $\mathcal{L}(V, W_r)$. The a-posteriori probability $P_\Lambda(V|X_r)$ is calculated using the log-linear combination of the feature functions based on the given model weights Λ as already introduced in Eq. (6.3):

$$P_\Lambda(V|X_r) = \frac{e^{\sum_i \lambda_i g_i(V,X_r)}}{\sum_{V'' \in \mathcal{M}_r} e^{\sum_i \lambda_i g_i(V'',X_r)}}, \tag{6.12}$$

where V is one of the concurrent word sequences $\mathcal{M}_r$ produced during recognition. Thus, $\mathcal{M}_r$ denotes the set of word sequences which are considered for discrimination in utterance r. In ideal case, $\mathcal{M}_r$ contains all possible sequences of the words contained in the recognition lexicon. In order to reduce computation time, a subset is chosen in practice which contains the best word sequences produced by the recognition system. For each utterance r, the set of concurrent word sequences is stored in the form of an N-best list. Similar to [Beyerlein 1998], N-best lists have been used in this thesis because no algorithm was available for the efficient calculation of the exact word error $\mathcal{L}(V, W_r)$ over word lattices. Recently, a new method has been proposed in [Heigold et al. 2005] which solves this problem allowing the use of word lattices in the MWE criterion. Nevertheless, in this work model weights have been trained using N-best lists.

In the following, a gradient descent method is formulated to find the optimal value of λ_i. The optimization process is defined as follows:

$$\lambda_i^{(j+1)} = \lambda_i^{(j)} - \epsilon \frac{\partial F(\Lambda^{(j)})}{\partial \lambda_i} \tag{6.13}$$

where the derivative of $F(\Lambda)$ with respect to λ_i is

$$\frac{\partial F(\Lambda)}{\partial \lambda_i} = \sum_{r=1}^{R} \sum_{V \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V|X_r)\ \eta\ g_i(V, X_r)\ \tilde{\mathcal{L}}(V, W_r, \Lambda) \tag{6.14}$$

and

$$P_{\Lambda \cdot \eta}(V|X_r) = \frac{e^{\sum_i (\eta\lambda_i) g_i(V,X_r)}}{\sum_{V'' \in \mathcal{M}_r} e^{\sum_i (\eta\lambda_i) g_i(V'',X_r)}} \tag{6.15}$$

$$\tilde{\mathcal{L}}(V, W_r, \Lambda) = \mathcal{L}(V, W_r) - \sum_{V' \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V'|X_r)\ \mathcal{L}(V', W_r). \tag{6.16}$$

Details of the derivation can be found in Annex B.1. The auxiliary quantity in Eq. (6.15) is a consequence of scaling the a-posteriori probability in Eq. (6.11) by η. In case of log-linear modeling, the scaling by η reduces to a simple multiplication of the model weights. Note that η does not need to be addressed explicitly in the implementation. Multiplication of the initial model weights by η automatically involves the scaling of optimized ones. Finally, note that the quantity $\tilde{\mathcal{L}}(V, W_r, \Lambda)$ defined in Eq. (6.16) can be interpreted as the number of word errors with "zero-mean".

The implementation of optimization procedure can be summarized in the following steps:

1. Generation of N-best lists $\mathcal{M}_r$ for the development corpus.
2. Calculation of feature function values $g_i(V, X_r)$ for every sentence V in each N-best list $\mathcal{M}_r$.
3. Initialization of model weights λ_i and of the a-posteriori probability scaling η.
4. Iteration of gradient descent optimization about $J = 100$ times.

The N-best lists are generated on the development corpus. The size of the N-best lists is determined in such a way that the N-best list error rate (cf. graph error rate) is half of the overall WER produced by the first-best recognition. In every N-best list and for each sentence, the feature function values are calculated using the corresponding standard rescoring algorithms, as summarized in Section 6.2.2.2. The rescored N-best lists are passed to the gradient descent algorithm. The initialization of the model weights and of the scaling of the a-posteriori probability η in Eq. (6.11) is carried out before the first iteration is started. Details on the initialization are presented in Section 6.2.2.3. In this work, instead of the application of a stopping criterion, fix amount of gradient descent iterations have been carried out. About $J = 100$ iterations have been abundantly enough in all experiments. The learning rate ϵ has been kept constant at a value of 0.001 throughout the feature combination experiments.

6.2.2.2 Calculation of Log-Linear Feature Values

The feature function values $-\log p_i(V, X_r)$ have been obtained by applying the corresponding standard rescoring algorithms to the N-best list. The language model and pronunciation model rescoring can be implemented in a straightforward way assuming N-best lists annotated with orthographies and pronunciation variants. Rescoring algorithms for emission and transition models operate on Viterbi alignments. For each word in the N-best lists, the start and end times are kept constant during the complete DMC training procedure. Alignments are generated separately for each word in each sentence using the standard Viterbi search algorithm. The HMM of the words considers the coarticulation at the word boundaries as well. Once the alignments are known, emission or transition scores can be summed up in a straightforward way for each word and finally for each sentence.

6.2.2.3 Initialization of Model Weights

The initialization of λ_i values has been done corpus dependent with empirical values typical for the model type. In the following typical values are mentioned for each feature function type:

- language model weight: $\lambda_{lm} = 1.0$

- pronunciation model weight: $\lambda_{pr} = 1.0$
- word penalty weight: $\lambda_{wrd} = 0.5$
- transition model weight: $\lambda_{tr_w} = 0.1$
- emission weights: $\sum_i \lambda_{f_i} = 0.05$

Besides the ratio of the initial weights, their absolute value plays a critical role during the DMC training. In Eq. (6.11), a scaling of the a-posteriori probability η is introduced explicitly to underline the influence of the absolute values of the initial model weights. Increasing η makes the a-posteriori probability Eq. (6.12) approximate the decision of the Bayes' classificator along with modifying the optimum criterion Eq. (6.11). The optimal value of η has been found empirically on the development corpus. On both large-vocabulary corpora, a value of $\eta = 3$ turned out to be a good compromise. A histogram of a-posteriori probabilities has been estimated to evaluate this case. The histogram shown on Figure 6.1 has been estimated on the first (i.e. recognized) sentence of the N-best lists generated for every utterance of the development set of the *EPPS* corpus. The a-posteriori probabilities have been calculated using the initial λ_i values multiplied by $\eta = 3$. The

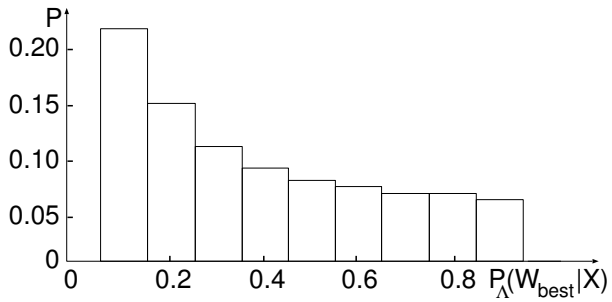


Figure 6.1: Histogram of a-posteriori probability Eq. (6.12) of the best sentence in the N-best lists for $\eta = 3$ on the *EPPS* development corpus.

histogram shows that despite the relatively high η the a-posteriori probabilities of the best sentences are significantly less than 1. (Note that in case of $\eta = \infty$, all a-posteriori probabilities are $\equiv 1$.) This indicates that the a-posteriori probabilities are still smoothed by suboptimal hypotheses in the N-best lists. With other words, suboptimal sentences in an N-best list $\mathcal{M}_r$ are included in the estimation of a-posteriori probabilities Eq. (6.12) with a significantly high weight. (In the extreme case $\eta = \infty$, only the best sentence in $\mathcal{M}$ determines the denominator.)

6.3 Integration of Log-Linear Model

The DMC recognition tests can be carried out in two different ways. The first option is a multi-pass method using word lattices to restrict the search space. The second possibility is the integration of the log-linear model combination into a standard search algorithm.

In the multi-pass method, word lattices are produced in the first step using the baseline recognition system. In the second pass, the word lattices are rescored by the different feature functions. The third pass consists of a search on the weighted sum of the rescored lattices yielding the recognized sentence.

The integration of the log-linear model into a standard search method facilitates a single-pass recognition. In this case, the weighting of the different knowledge sources happens on demand. Except the emission feature function, a single instance of the knowledge source has been used in our experiments. The on-demand weighing of the language, pronunciation, and transition feature functions can be implemented in a straightforward way. The score delivered by these knowledge sources is simply multiplied by the corresponding model weight before passing the value to the parent algorithm.

The integration of the emission feature functions requires more efforts. As described in Section 6.2, for each acoustic feature a separate emission model is trained consisting of a set of GMMs and a Classification and Regression Tree (CART). The CART tree provides a clustering of triphone states to reduce the number of Gaussian mixture distributions. The GMMs produce the emission probabilities for each class in the CART tree. Construction of data structures for the recognition process requires the number of triphone classes in advance. Since the combination of emission models containing different state tying trees results in a 2-tuple of triphone state class indexes, the amount of possible combination has to be determined before the recognition can be started. The maximum number of possible tuples is $|t_1| \cdot |t_2|$ where the CART trees are denoted by t_1 and t_2 and $|\cdot|$ provides the number of classes differentiated by a single tree. However, redundancy in acoustic features leads to redundancy in the state tying trees as well. Consequently, the real number of class index tuples is significantly less. In the following, an algorithm is presented which combines two CART trees t_1 and t_2 resulting in tree t whose leaves do not contain class indexes anymore but 2-tuples of indexes:

```
# copy the t_1 to create t
t = copy(t_1)

# copy the original set of leaves of t since t will be changed
L = leaves(t)

FOR EACH l in L:
    # create temporary tree t' by copying t_2
    t' = copy(t_2);

    # collect all {question, answer} pairs in Q from l backward to the root of t
    Q = {};
    k = l;
    WHILE k ! = root(t):
        p = parent(k);
        Q = {Q, {question(p), answer(p, k)}};
        k = parent(k);

    # replace nodes with the appropriate child if answer known
    FOR EACH {q,a} in Q:
        WHILE [n = find_node(t', q)] != null:
            n := child(n, a);

    # extend all leave index tuple in t' by the index of l
    FOR EACH l' in leaves(t'):
        index(l') := {index(l), index(l')};

    # overwrite l by the root of t'
    l := root(t');
```

In the case of three or more models, the algorithm can be applied successively. In Section 7.1.1, recognition experiments are presented combining two or three emission models. The increasing number of class index tuples shows that the log-linear model introduces a much larger variability in the amount of Gaussian mixtures when compared to single-feature systems.

The combination of emission scores happens on-the-fly. Different acoustic feature vectors sharing the same time frame index t are bunched and the tuple is fed into the recognition algorithm. The emission feature functions (i.e. emission GMM sets) of different acoustic features are kept parallel in the memory. In every time frame, each emission feature function receives the corresponding feature vector. Each acoustic feature is evaluated by the corresponding GMM for every triphone state given by the search algorithm. Output scores of GMMs are weighted by the corresponding model weight λ_{f_i}. Finally, the sum of the weighted scores is passed back to the search algorithm as emission score.

6.4 Recognition Tests

In this work, experiments have been carried out by using the integrated weighting, i.e. integration of the log-linear model into the standard search algorithm. Nevertheless, combining a higher number of emission models increases the computational costs so much so that the application of the rescoring methods becomes inevitable. Table 6.1 summarizes the recognition results obtained by combining the acoustic features introduced in Chapter 4.

The DMC based combination of different acoustic features denoted by Σ_{DMC} has resulted

Table 6.1: Word error rate obtained by applying discriminative model combination (DMC) to single and LDA-combined features.

acoustic feature	error rates [%] del	ins	WER
VM II corpus			
MFCC	4.5	2.9	21.0
MF-PLP	5.2	2.3	21.0
PLP	5.9	2.3	21.4
Σ_{DMC}	4.7	2.7	20.4
MFCC	4.5	2.9	21.0
V	n/a	n/a	n/a
Σ_{DMC}	5.3	2.5	21.2
LDA(MFCC+V)	4.6	2.7	20.3
LDA(VTLN+V)	4.1	2.7	18.7
Σ_{DMC}	4.1	2.5	18.4
EPPS corpus			
MFCC	4.3/3.8	1.4/1.7	14.7/15.3
MF-PLP	4.2/3.7	1.5/1.7	14.8/15.3
PLP	4.3/3.5	1.6/1.8	15.4/15.8
Σ_{DMC}	4.4/3.8	1.3/1.4	14.5/14.5
MFCC	4.3/3.8	1.4/1.7	14.7/15.3
V	n/a	n/a	n/a
Σ_{DMC}	4.2/3.7	1.4/1.7	14.8/15.2
LDA(MFCC+V)	3.9/3.4	1.5/1.9	14.3/14.8
LDA(VTLN+V)	4.0/3.3	1.5/1.6	13.8/14.0
Σ_{DMC}	4.1/3.5	1.2/1.4	13.6/13.5

either in significant improvements or in insignificant changes in WER relative to the result obtained by using the best performing feature among the combined ones. The combination of the baseline acoustic features MFCC, MF-PLP, and PLP resulted in significant improvements in WER of up to 5% relative to the using the MFCC feature alone. This result corresposponds to observations published in [Häb-Umbach and Loog 1999].

The DMC based combination of the MFCCs with the voicing feature (V) did not yield

any significant changes. The acoustic model of the voicing measure has been trained with using the standard maximum likelihood training. Note that Viterbi alignments have not been generated since this feature does not carry sufficient amount of information. An alignment generated by using the MFCC feature has been used. Several experiments have been run to optimize the size of the CART tree and the number of Gaussian densities. Despite the extensive optimization, the combination of the MFCC and the voicing features has not improved the WER.
Results obtained by using the LDA based combination can be improved by nesting the two combination methods. Features combined by using LDA can be interpreted as a new separate acoustic feature. The emission models trained on these features can be combined again using the DMC framework. For example, nesting the LDA combined features LDA(MFCC+V) and LDA(VTLN+V) into the DMC based method has yielded significant improvements in WER on both tested corpora.

Table 6.2 demonstrates the weighting of different knowledge sources estimated by the MWE training. The weights have been obtained in the experiments presented in Table 6.1. The model weights have been divided by the language model weight λ_{lm} to provide comparable values. The emission feature function weights λ_{f_i} play the most important role when combining acoustic features. However, changes in the emission modeling are reflected in the within-word transition λ_{tr_w}, word penalty λ_{wrd}, and pronunciation λ_{pr} weights as well. The emission weights provide information on the contribution of the

Table 6.2: Model weights obtained by applying discriminative model combination to single and LDA-combined features. (All model weights are divided by the language model weight.)

corpus	acoustic features	model weight					
		λ_{f_i}			λ_{tr_w}	λ_{wrd}	λ_{pr}
VM II	MFCC+MF-PLP+PLP	0.018	0.017	0.018	0.103	0.499	n/a
	MFCC+V	0.045	0.024		0.098	0.500	n/a
	LDA(MFCC+V)+ LDA(VTLN+V)	0.013	0.038		0.093	0.499	n/a
EPPS	MFCC+MF-PLP+PLP	0.025	0.014	0.013	0.009	0.497	0.995
	MFCC+V	0.054	0.046		0.000	0.486	0.974
	LDA(MFCC+V)+ LDA(VTLN+V)	0.015	0.036		0.019	0.499	0.999

different acoustic features to the final emission score. The values cannot be compared directly without respect to the ranges of the underlying feature function outputs. However, the model weights of the MFCC, MF-PLP, and PLP features are comparable since they yield emission scores in the same range. On the *VM II* corpus, all the three features have received similar weights. The PLP feature is slightly suppressed in the same setup on the *EPPS* corpus which can be expected looking at its recognition performance. Nevertheless, even the worst performing feature has a contribution to the improved final recognition results. The same observation can be done when combining the two

LDA combined features. The LDA(VTLN+V) and the LDA(MFCC+V) features output emission scores in the same range as well. Despite its much weaker recognition results, the LDA(MFCC+V) feature contributes approximately one third of the total emission score yielding improvement in WER compared to using the LDA(VTLN+V) feature alone. The model weights obtained by the combination of the MFCC and the voicing (V) features cannot be compared directly. The ratio of the average emission scores of the two features is ≈ 40. Despite the small scores produced by the voicing emission model, the discriminative training did not increase the contribution of this feature. Even manually increased weights have been tested to increase the impact of the voicing emission model. The baseline recognition results could not be improved either. The success of the LDA based combination of this feature pair ensures that the voicing feature contains classification information additional to the MFCC feature. A successful way of the integration of the voicing feature into the DMC framework has not been found in the scope of this thesis. Further analysis of this problem may even contribute to a better understanding of attributes of DMC.
Besides the emission weights, weights of the standard knowledge sources are adjusted automatically by the MWE training. The transition model within word boundaries has played a much more important role on the *VM II* than on the *EPPS* corpus. The differences are due to the optimization which have been performed extra on the *VM II* corpus. The word penalty plays approximately the same role on both corpora. The pronunciation model has only been applied to the *EPPS* corpus. Both model weights stayed almost unchanged compared to their initial values. A probable explanation is provided by the empirically optimization of these knowledge source weights. The WER curves are rather flat around the optimum weight. Therefore a relatively good initial value yields already the best recognition results.

6.5 Summary

In this chapter, an application of the Discriminative Model Combination has been introduced to combine acoustic features. After the introduction of the log-linear model and the corresponding standard feature functions, a splitting of the acoustic model has been derived to achieve separately scalable acoustic knowledge sources. It has been shown that the splitting requires the use of maximum approximation in place of the summation over all pronunciation variant and all HMM state sequences during the decoding process. A general training of log-linear model weights has been described using the Minimum Word Error criterion. Recognition results have been presented by using the combination of auditory and articulatory motivated features. The conclusions of recognition tests can be summarized as follows:

- According to expectations, no acoustic feature set has been found which has led to degradation in WER compared to the best performing feature in the set. The MWE training ensures that model weights are optimized with respect to the WER. The training suppresses acoustic features with a low log-linear weight which cannot contribute to the improvement of the WER.

- A surprising observation is that the combination of rather similar features such as the MFCC, PLP, and MF-PLP features can lead to significant improvements in WER of up to 5% relative to using solely the best performing feature.
- The DMC based combination of LDA combined features can be implemented in a straightforward way. Nesting the two methods yields improvements in WER of up to 3% relative to the best performing LDA combined setup.

Chapter 7

Comparison of LDA and DMC Based Combinations

Although the LDA based and the DMC based feature combination techniques operate at different levels of speech recognition systems, the resulting combined emission models share a common structure. In spite of a few empirical results, a comprehensive comparison of these feature combination techniques has not been published yet. One of the first empirical comparisons of these methods was performed in [Häb-Umbach and Loog 1999]. Experimental results have been presented on the combination of three standard cepstral acoustic features. The two feature combination methods considered in this thesis have been extended by a method called ROVER which combines recognition systems on the level of recognized words. Recognition results have shown that using the DMC based method yields the most improvement in WER compared to the best single-feature system. In [Zolnay et al. 2005], a comparison of the LDA and log-linear model combination was first presented using not only different cepstral features but also the one-dimensional voicing measure. The best recognition results were obtained by nesting the LDA based method into the log-linear model combination. Note that in this paper, the log-linear model weights were estimated by using a simple grid search which requires careful manual adjustment of the grid points.

In this Chapter, the goal of the comparison is to point out drawbacks of the LDA and DMC based methods rather than to find the better combination technique since the two methods can be nested in a straightforward way. A theoretical comparison of the LDA and the DMC based methods is presented in Section 7.1 which considers the similarities and differences of the underlying emission models. In Section 7.2, a review of recognition results is presented which have been obtained by applying either LDA or DMC to the same set of acoustic features. The recognition results are compared and discussed with respect to the conclusions of the theoretical comparison.

7.1 Theoretical Comparison

In order to find the major differences between the LDA and DMC based methods, underlying emission models are compared with respect to their structure and to the number of parameters involved in the combination. Both combination methods include the LDA and the log-linear modeling. The approaches are differentiated by the choice in which of the two algorithms the actual combination takes place. The LDA based approach combines the feature vectors directly followed by the training of a single emission model. The log-linear model is in this case only responsible for the combination of the emission model with the rest of the knowledge sources introduced in Eq. (6.10). When applying DMC, an LDA matrix is estimated for every feature separately which finds the best linear combination of successive feature vectors. In this case, the feature combination takes place in the log-linear model. A discriminative training estimates the weights of the different features along with the weights of the rest of the knowledge sources.

In case of the LDA based combination, the log-linear model combination includes a single emission probability whereas the DMC based method results in the product of scaled emission probabilities, as can be seen in the last term of Eq. (6.10). In the RWTH recognition system, Gaussian Mixture Models (GMM) are used to model emission distributions. For the sake of simplicity, in the following, the product of scaled Gauss distributions is analyzed rather than the scaled product of GMMs. Note that using maximum approximation in GMMs reduces the scaled product to this simple case. Hence, the results of the following derivation can be applied directly to GMMs. The product of scaled Gauss distributions can almost be transformed into a single multidimensional Gauss distribution. As the example of two feature extraction algorithms x and y shows in Eq. (7.1a), the resulting distribution is almost a Gauss one with a block diagonal covariance matrix. The underlying covariance matrices Σ_x and Σ_y are located in the diagonal, divided by the model weights λ_x and λ_y, respectively:

$$\mathcal{N}(x|\mu_{\phi_x(s)}\ \Sigma_x)^{\lambda_x} \cdot \mathcal{N}(y|\mu_{\phi_y(s)}\ \Sigma_y)^{\lambda_y} = \tag{7.1a}$$

$$\sqrt{\lambda_x^{dim(x)}|\Sigma_x|^{(\lambda_x-1)}\lambda_y^{dim(y)}|\Sigma_y|^{(\lambda_y-1)}}^{-1} \cdot \tag{7.1b}$$

$$\mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix} \middle| \begin{bmatrix} \mu_{\phi_x(s)} \\ \mu_{\phi_y(s)} \end{bmatrix} \begin{bmatrix} \frac{1}{\lambda_x}\Sigma_x & 0 \\ 0 & \frac{1}{\lambda_y}\Sigma_y \end{bmatrix}\right) \tag{7.1c}$$

As expected, the distribution is not normalized anymore. The difference is caused by the multiplicative term Eq. (7.1b) depending on the covariance matrices and the model weights. Nevertheless, this term can be neglected under the assumption of globally pooled covariances. As (7.1a) reflects, the assumption of globally pooled covariances per acoustic feature holds in the scope of this thesis. Hence, the term Eq. (7.1b) is constant over all states s and it can be excluded from the emission score calculation. Accordingly, a comparison can consider the scaled product of Gauss distributions as a single Gauss distribution in the form shown in Eq. (7.1c).

In the following, the single GMM set based emission model used in the LDA combined systems is compared to the combined GMM sets obtained by using log-linear combination. The two emission models are first compared in terms of the triphone state tying

$\phi_{f_i}(s)$ structure in Section 7.1.1 followed by the analysis of the amount of combination parameters in Section 7.1.2.

7.1.1 State Tying Structure

In the LDA based system, a single emission GMM set is trained using a single Classification and Regression Tree (CART) for state tying. In the large-vocabulary experiments, GMMs have been split seven times resulting in $\approx$ 400k mean vectors depending on the corpus, see Table A.1. In the DMC based feature combination, a separate GMM set is estimated for each acoustic feature. Consequently, the number of mean vectors is $\approx$ 400k times the number of different acoustic features. Besides the fairly increased number of mean vectors, the DMC based method changes the number of CART classes i.e. the number of triphone state classes, too. Instead of using a single global CART tree, a separate CART tree $\phi_{f_i}(s)$ has been estimated for each acoustic feature f_i. As shown in the example in Eq. (7.1a), the separate trees $\phi_x(s)$ and $\phi_y(s)$ result in a 2-tuple of class indexes for every triphone state s. Theoretically, the maximum number of index tuples is $\prod_i |\phi_{f_i}(s)|$ where $|\cdot|$ denotes the number of classes of a single tree. In practice, the redundancy between the acoustic features leads to much lower number of active index tuples. Using the CART tree combination algorithm presented in Section 6.3, the real number of index tuples can be determined before the recognition process is started. Table 7.1 compares the number of CART classes respectively CART index tuples used by the two feature combination methods. Hence, the two methods do not only differ

Table 7.1: Number of CART classes or CART index tuples obtained when combining the same acoustic features with using the LDA and DMC based feature combination methods.

corpus	acoustic features	# CART index tuples
VM II	LDA(MFCC+VTLN+V)	3 501
	DMC[(LDA(MFCC+V)+ LDA(VTLN+V)]	90 128
	LDA(MFCC+MF-PLP+PLP)	3 501
	DMC(MFCC+MF-PLP+PLP)	604 157
EPPS	LDA(MFCC+VTLN+V)	4 501
	DMC[(LDA(MFCC+V)+ LDA(VTLN+V)]	174 143
	LDA(MFCC+MF-PLP+PLP)	4 501
	DMC(MFCC+MF-PLP+PLP)	655 223

by the total amount of mean vectors but also in the number of mean vector combinations induced by the feature dependent state tying when using DMC.

7.1.2 Amount of Combination Parameters

In this section, the two feature combination methods are compared in terms of the amount of parameters explicitly controlling the feature combination. The LDA based method uses much more combination parameters than the DMC based one. On both of the tested large-vocabulary corpora, the LDA based method generates 45 feature components by the linear combination of successive vectors generated by different extraction methods. Depending on the dimensionality of the features and the LDA window length, the number of parameters directly combining the features is approximately 10^4. The DMC based method operates with much less parameters. Despite the implicit effects of the emission model parameters, only the few emission model weights control the influence of the combined features. The number of these weights is equal to the number of different acoustic features which hardly grows up to 10 in our current experimental setups.

Furthermore, there is one more significant difference between the two sets of combination parameters. The DMC based method estimates the model weights by using a discriminative criterion. Whereas, it has been shown, that LDA can also be derived from a maximum likelihood criterion which is not directly bound to word error rate.

Finally, a DMC based feature combination experiment is recalled in which the MFCC and the voicing features have been combined. The large difference in the amount of combination parameters may explain why the LDA based combination outperforms the DMC based approach. The LDA matrix operates directly on several successive time frames of the two feature streams whereas DMC has only access to two acoustic scores per time frame as classification information. This indicates that one of the straightforward extensions of the presented feature combination systems would be discriminative training of LDA matrices applied to acoustic feature combination.

7.2 Comparison of Recognition Results

In this section, experimental results are reviewed which have already been reported in Chapter 5 and 6. The experiments performed using the LDA and DMC based feature combination methods are presented in pairs, i.e. the same acoustic features combined by using the two methods. The goal is to locate setups where the advantages or disadvantages of each method become visible.

Table 7.2 presents the experiment pairs performed on the *VM II* and *EPPS* corpora. The recognition results can be interpreted as follows:

1. The first two experiments define the baselines without and with speaker normalization (i.e. Vocal Tract Length Normalization).

2. The second group of experiments have combined the MFCC, MF-PLP, and PLP features. It is obvious that in this case the DMC based combination method is superior to the LDA based one which is most probably inhibited by either numerical instabilities or insufficient amount of training data. The DMC based combination

Table 7.2: Word error rates obtained by applying Discriminative Model Combination (DMC) based and Linear Discriminant Analysis (LDA) to the same set of features.

corpus	acoustic features	error rates [%] del	ins	WER
VM II	MFCC	4.5	2.9	21.0
	VTLN	3.8	2.9	19.1
	LDA(MFCC+MF-PLP+PLP)	4.7	3.3	21.6
	DMC(MFCC+MF-PLP+PLP)	4.7	2.7	20.4
	LDA(MFCC+V)	4.6	2.7	20.3
	DMC(MFCC+V)	5.3	2.5	21.2
	LDA(MFCC+VTLN+V)	3.5	2.9	18.2
	DMC[(LDA(MFCC+V)+LDA(VTLN+V)]	4.1	2.5	18.4
EPPS	MFCC	4.3/3.8	1.4/1.7	14.7/15.3
	VTLN	4.3/3.7	1.3/1.5	14.2/14.1
	LDA(MFCC+MF-PLP+PLP)	4.8/4.1	1.4/1.6	15.8/16.2
	DMC(MFCC+MF-PLP+PLP)	4.4/3.8	1.3/1.4	14.5/14.5
	LDA(MFCC+V)	3.9/3.4	1.5/1.9	14.3/14.8
	DMC(MFCC+V)	4.2/3.7	1.4/1.7	14.8/15.2
	LDA(MFCC+VTLN+V)	4.0/3.5	1.5/1.6	13.7/14.1
	DMC[(LDA(MFCC+V)+LDA(VTLN+V)]	4.1/3.5	1.2/1.4	13.6/13.5

has yielded improvements in WER of up to 5% relative to using the MFCC feature alone.

3. In the third group of experiments, the MFCC feature has been combined with a voicing measure (V). The LDA based method has improved the WER by 3% relative to using the MFCCs alone whereas the DMC based combination of the two feautes has not yielded any significant change in WER. Despite intensive optimization of the CART tree and the emsission GMM set trained on the voicing measure, the DMC based combination could not improve the baseline recognition results. Further investigations are necessary on this feature pair before a conclusion can be drawn.

4. The fourth group of experiments combines the MFCC, VTLN, and voicing features. Since the DMC based combination of the MFCC and voicing features has turned out to be unsuccessful, an alternative way has been chosen to test the DMC based combination of the three features. Hence, the acoustic models trained on the LDA(MFCC+V) and LDA(VTLN+V) features have been combined again using DMC. On the *VM II* corpus, the DMC based combination has resulted in a small improvement in WER compared to the LDA(VTLN+V) system, cf. Table 6.1. However, the pure LDA based combination of the three features performed slightly better than the DMC system. Improvements in WER of up to 5% have been obtained relative to using the best performing baseline system (VTLN). Different results have been observed on the *EPPS* corpus. The DMC experiment has yielded fair improvements compared either to the LDA(VTLN+V) or to the

LDA(MFCC+VTLN+V) systems. The combination of the three features has resulted in improvements of up to 4% relative to using the best baseline system.

7.3 Summary

In this chapter, a comprehensive comparison of two feature combination methods has been presented. The similarities and differences between the LDA and DMC based methods can be summed up as follows:

- Both combination methods have yielded consistent results on both large-vocabulary corpora. Improvements in WER of up to 5% have been obtained by using combined features compared to the speaker adapted single-feature systems.
- Advantageous feature combination setups have been found for both methods, i.e. none of the methods has been found to be obviously superior.
- The DMC based method produces a much higher variability in the emission model by combining feature dependent CART trees.
- The number of parameters explicitly involved in the feature combination is with orders of magnitude larger when using LDA.
- The LDA based method operates directly on feature vectors of successive time frames produced by the different feature extraction algorithms whereas DMC receives classification information reduced into single acoustic scores.
- The training procedure of LDA can be derived from a maximum likelihood criterion wheres the DMC training is directly bound to the word error rate.
- The LDA based feature combination method is the more likely to fail, the more correlated the combined features are. Therefore, a careful selection of the features or feature vector components is required. The DMC based method is robust against dependencies between acoustic features and against for the recognition irrelevant features also due to the discriminative training.
- Consistent improvements have been obtained by applying DMC to feature combination. The LDA based combination of acoustic features can be nested into the DMC method yielding the best recognition results. The most remarkable improvements have been obtained on the evaluation part of the *EPPS* corpus. On this corpus, the DMC based feature combination has yielded an improvement in WER of 4% relative to the best LDA based system.

Chapter 8

Scientific Contributions

The aim of this thesis was twofold: On the one hand, new articulatory motivated features were developed and tested in combination with state-of-the-art features. On the other hand, acoustic feature combination techniques were described and compared at a theoretical and a practical level. In particular, the following aspects were studied:

Voicing measure

Besides considering four state-of-the-art baseline acoustic features, the extraction of the voicing feature was investigated. The voicing feature captures the level of vibration of the vocal cords in the form of a bounded scalar value. Three partially new voicing extraction techniques were studied. The experimental comparison did not show any significant difference in the performance of the three methods. The pairwise theoretical comparison revealed common basic structures, however, significant differences were found in the normalization of the voicing measures and in the computational costs. The *autocorrelation* based voicing measure was chosen because of its practically most attractive properties.The voicing measures were tested in combination with several baseline features. On the small-vocabulary *SieTill* task, the LDA based combination of the MFCC feature with a voicing measure resulted in an improvement in WER of 11% relative to using the MFCC feature alone. Smaller but consistent improvements were obtained on the large-vocabulary *VM II* and *EPPS* tasks. The combination of the voicing measure with one of the baseline features yielded improvements of up to 3% relative to using the baseline feature alone.

Spectrum derivative measure

The spectrum derivative feature focuses on the differentiation of obstruent from sonant consonants by summarizing the intensity of changes of the magnitude spectrum in the form of a scalar value. The feature is a product of recently research still exhibiting a few unclear properties. Nevertheless, experiments carried out in combination with state-of-the-art speaker independent and dependent baseline features resulted in significant improvements in WER. On the small-vocabulary *SieTill* task, the LDA based combination of the MFCC and the spectrum derivative features yielded an improvement in WER similar to using the voicing feature. An improvement of 11% was obtained relative to using the MFCC feature alone. Smaller and inconsistent improvements were found on the large-

vocabukary *VM II* and *EPPS* tasks. The additional spectrum derivative feature yielded improvements of up to 3% on the *VM II* corpus, however, no significant improvement was found on the *EPPS* corpus.

Linear Discriminant Analysis

A detailed analysis of the LDA based combination method was presented which addressed problems published formerly along with difficulties arising from its application to feature combination. Different experiments were conducted to test the numerical robustness of LDA against increasing input dimension. The reciprocal condition number of eigenvalues and eigenvectors was used to determine the reliability of the estimated projector matrices. The conclusions can been summarized as follows:

- The LDA based combination of increasing number of acoustic features can improve the WER monotonously. Experiments on the small-vocabulary *SieTill* task have yielded a relative improvement in WER of 16% when adding the voicing and the spectrum derivative features to the MFCCs. Experiments conducted on large-vocabulary tasks have resulted in monotonous improvements in WER as well. Improvements of up to 4% have been obtained relative to the best performining single feature systems. However, a set of features (MFCC, MF-PLP, and PLP) was found whose combination yielded significant degradation in WER even compared to the worst performing feature in the set. Conclusions of the analysis of this case are presented in the following bullet points.
- Experiments with additive white noise components indicate that the input dimension of LDA is not the bottleneck of the method. An input of ≈ 1000 components leads still to a stable estimation of the projection matrices.
- Stronger correlation between the feature vector components causes ill-conditioned eigenvalue problem leading to unstable estimates of the projection matrix and consequently to degradation in WER. Standard algorithms for solving the generalized eigenproblem do not address the case of almost singular matrices explicitly leading to unreliable eigenvalue estimates. Special algorithms developed for general singular pencils [Demmel and Kagström 1993a][Demmel and Kagström 1993b] or for symmetric singular pencils [Parlett and Chen 1990] should be considered to improve the eigenvalue and eigenvector estimates.
- The analysis of numerical stability by means of the reciprocal condition numbers could not explain the degradation in WER when increasing the LDA window length and when combining the MFCC, MF-PLP, and PLP features. Hence, the sparseness of the scatter matrix estimates remains one of the most probable explanations.

Discriminative Model Combination

An application of DMC was presented for feature combination. A splitting of the acoustic model was carried out resulting in independent log-linear feature functions. A separate emission feature function was introduced for each acoustic feature. The feature function weights were trained by optimizing the Minimum Word Error criterion. Recognition tests were carried out on the two large-vocabulary corpora. The conclusions are the following:

- Despite their common basic structure, the combination of the baseline MFCC, MF-PLP, and PLP features resulted in significant improvements in WER of up to 5% compared to the best performing MFCC feature.
- Unlike using LDA, the DMC based combination of the single voicing measure with the MFCC feature resulted in no changes in WER.
- The best recognition results were obtained by using the voicing measure in combination with the MFCC and VTLN features. The LDA based combination of these features nested into the DMC yielded improvement in WER of up to 3% over the best underlying LDA system.

Comparison of LDA and DMC Based Combination Methods

Finally, a comparison of the two feature combination methods was performed. In the theoretical comparison, the log-linear combination of emission models was reduced to a single emission model. The comparison of the emission models trained on the LDA combined features with the reduced log-linear emission model showed that there are major differences between the two methods:

- The DMC methods produces a much higher number of Gauss densities.
- The number of parameters explicitly involved in the feature combination is with orders of magnitude larger when using LDA.

The practical comparison of the methods showed the following:

- The DMC based method is robust against dependencies between combined features whereas LDA based combination of such features leads even to degradation in WER.
- Advantageous feature combination setups were found for both methods, i.e. none of the methods was found to be obviously superior.
- The best recognition results were obtained by nesting the LDA based combination into the DMC. A consistent improvement in WER of up to 4% relative to the baseline VTLN system was obtained on both corpora.

Chapter 9

Outlook

In this thesis, several acoustic features and their combination was studied. It was shown that articulatory information can improve the accuracy of the state-of-the-art speech recognition systems. However, the new acoustic features leaded to improvements in WER only when appropriately combined with the baseline features. The LDA and the DMC based feature combination methods were used in feature combination experiments. The theoretical and experimental comparison of the methods did not find an obviously superior feature combination technique. Advantageous feature setups were found for both method. The best recognition results were obtained by nesting the two methods. The following questions remain open and may serve as starting point for further research:

Acoustic Feature Extraction

- Simple acoustic features were developed in this thesis which focus on a single articulatory class. The combination of these articulatory cues leaded to significant improvements in WER. Hence, the development of features distinguishing articulatory classes could lead to further improvements in accuracy of automatic speech recognition systems.

- Alternative feature extraction techniques are emerging such as pith-synchronous signal analysis, phase spectrum based acoustic features, etc.

- The optimization of the existing features may draw attention. The voicing extraction is a topic of ongoing research. The optimal integration of the newest voiced-unvoiced detection algorithms into the speech recognition remains an interesting topic. Promising recognition results were obtained by using the additional spectrum derivative feature. However the results are not consistent over all corpora. Further analysis of the spectrum derivative feature can become a subject of future research.

Linear Discriminant Analysis

- The analysis of reciprocal condition numbers showed that stronger correlation between combined feature vector coefficients leads to ill-conditioned eigenproblem. State-of-the-art algorithms do not consider

this case. Special algorithms developed for general singular pencils [Demmel and Kagström 1993a][Demmel and Kagström 1993b] or for symmetric singular pencils [Parlett and Chen 1990] should be considered to improve the eigenvalue and eigenvector estimates.

- The analysis of the numerical stability of eigenvalue estimates could not give a satisfactory answer to all open problems. The reason of the degradation in WER when increasing the window length was not found either. The hypothesis of the sparseness of the scatter matrix estimates remains one of the most probable explanations. Experiments with increasing amount of training data may verify this hypothesis.

- It has been shown, that LDA can also be derived from a maximum likelihood criterion which is not directly bound to word error rate. This indicates that one of the straightforward extensions of the presented feature combination systems would be the discriminative training of LDA matrices. Although preliminary experiments on single-feature systems did not improve the WER, the application of discriminative training to LDA matrices may become beneficial when used for acoustic feature combination.

Discriminative Model Combination

- Recognition results obtained by using the RWTH Discriminative Model Combination toolkit were first published in this thesis. In these experiments, the log-linear combination of the underlying knowledge sources was integrated into a word-conditioned tree search algorithm allowing a single pass recognition. Increasing the number of acoustic or other sort of knowledge sources requires the implementation and the optimization of a word lattice rescoring based recognition system. The optimal structure and size of word lattices have to be found to utilize the sometimes slight differences between the knowledge sources.

- In this thesis, the Minimum Word Error optimum criterion was applied to the estimation of the model weights. Alternative optimum criteria are well known and are successfully applied to the discriminative training of acoustic models. The application of these criteria to DMC or the development of new criteria may become the focus of future research activities.

- The LDA based combination of the MFCC and the voicing features outperformed the DMC based one. In fact, the DMC based combination did not yield any significant change in WER. Despite intensive optimization of the acoustic model trained on the one-dimensional voicing measure, the WER could not be improved by applying the DMC. The reasons are not obvious. Searching them may provide a better insight into the properties of DMC.

Appendix A

Corpora and Recognition Systems

This annex summarizes information about the different speech corpora and about the different systems for small- and large-vocabulary speech recognition. The three distinct recognition tasks and the speech recognition systems used for the experiments are presented first. The chapter is concluded by a comparison of the different recognition scenarios in Table A.1.

A.1 SieTill

The small-vocabulary tests were performed on the *SieTill* corpus. The corpus consists of German continuous digit strings recorded over telephone line. The vocabulary includes the ten digits plus one pronunciation variant 'zwo' as variant for 'zwei'. Recognition tests have only been performed on the test set. The recognition system is based on a one-pass decoder design. Details on acoustic modeling are summarized in the following:

- telephone line recorded German digits;
- training corpus 11.6h, test corpus 11.7h;
- 11 whole word HMMs incl. 'zwo';
- per gender 214 states plus 1 for silence;
- HMM segments with 2 identical emission distributions;
- Gaussian mixture densities;
- pooled diagonal covariances;
- within-word acoustic model;
- 12 MFCC coefficients;
- LDA on 11 adjacent input frames which are reduced to 30 output coefficients.

A.2 VerbMobil II

The *VerbMobil II* (*VM II*) corpus consists of German human-to-human dialogs recorded in clean conditions. Recognition tests have only been performed on the test set. The recognition system is based on a one-pass word-conditioned tree search based decoder design. Details on acoustic modeling are summarized in the following:

- German large-vocabulary conversational speech;
- vocabulary: 10157 words with pronunciation variants and including spellings, hesitations, and noises;
- class-trigram language model ($PP_{test} = 62.0$);
- 3501 mixtures with a total of $\approx$ 396k Gaussian densities;
- one pooled diagonal covariance;
- across-word acoustic model;
- 3-state HMMs with skip;
- 16 MFCC coefficients;
- LDA on 11 adjacent input frames which are reduced to 45 output coefficients.

A.3 European Parliament Plenary Sessions

The European Parliament Plenary Sessions (*EPPS*) corpus from the 2005 TC-STAR ASR evaluation campaign contains plenary session speeches of the European Parliament in British English. The speeches are spoken either directly by politicians or by interpreters translated into English. Recognition result are presented on both the development and the evaluation sets. In tables, recognition results obtained on the two sets are shown successively in the following form: dev/eval. The recognition system is based on a one-pass word-conditioned tree search based decoder design. Details on acoustic modeling are summarized in the following:

- English large-vocabulary recognition of parliament plenary sessions;
- vocabulary: 54265 words;
- trigram language model ($PP_{dev} = 87.0$, $PP_{eval} = 99.0$);
- 4501 mixtures with a total of $\approx$ 446k Gaussian densities;
- one pooled diagonal covariance;
- across-word acoustic model;

- 6-state HMMs;
- 16 MFCC coefficients;
- LDA on 9 adjacent input frames which are reduced to 45 output coefficients.

A.4 Summary

Table A.1 compares the three corpora on which recognition experiments were performed within the scope of this thesis. Besides the typical properties of corpora, characteristical settings are presented which describe the corresponding recognitions systems.

Table A.1: Settings of the RWTH recognition system for the *SieTill*, *VM II*, and *EPPS* corpora.

corpora	name	*SieTill*		*VM II*		*EPPS*		
		train	test	train	test	train	dev	eval
	speech seg. [h]	11.6	11.7	61.5	1.6	40.8	3.7	3.5
	silence [%]	≈55	≈55	13	11	16	13	14
	# speakers	362	356	857	16	154	16	36
	# sentences	12 948	13 114	36 015	1 081	21 497	1 400	1 449
	# run. words	42k	43k	701k	14k	358k	34k	33k
lexicon	vocabulary size	11		10 157		54 265		
		train	test	train	test	train	dev	eval
	OOV [%]	n/a	n/a	n/a	n/a	0.7	1.1	2.6
	broken words [%]	n/a	n/a	n/a	n/a	0.6	0.8	2.1
language model	type	zerogram		class-trigram		trigram		
			test		test		dev	eval
	perplexity		n/a		62.0		87	99
feature extraction	sample rate [kHz]	8		16		16		
	# cepstrum	12		16		16		
	LDA window	11		11		9		
	dim. after LDA	30		45		45		
sub-word units	type	whole-word		triphone		triphone		
	gender dep.	yes		no		no		
	across-word	no		yes		yes		
HMM topology	(sub)words	≈ 40 state		3 states w/ skip		6 states		
	silence	1 state		1 state		1 state		
state tying	type	none		decision tree		decision tree		
	# GMM	215		3 501		4 501		
emission modeling	# densities	≈7k		≈396k		≈446k		
	pooled covar.	yes		yes		yes		
	diagonal covar.	yes		yes		yes		

Appendix B

Detailed Calculations

B.1 Derivation of MWE criterion

In this section, derivation of the Minimum Word Error (MWE) criterion is presented in detail. The MWE criterion is defined as:

$$F(\Lambda) = \sum_{r=1}^{R} \frac{\sum\limits_{V \in \mathcal{M}_r} [P_\Lambda(V|X_r)]^\eta \;\; \mathcal{L}(V, W_r)}{\sum\limits_{V' \in \mathcal{M}_r} [P_\Lambda(V'|X_r)]^\eta}. \tag{B.1}$$

In this particular case, the a-posteriori distribution is modeled by log-linear distribution:

$$P_\Lambda(V|X_r) = \frac{e^{\sum\limits_i \lambda_i g_i(V, X_r)}}{\sum\limits_{V'' \in \mathcal{M}_r} e^{\sum\limits_i \lambda_i g_i(V'', X_r)}}. \tag{B.2}$$

Inserting the log-linear model into the MWE criterion yields:

$$F(\Lambda) = \sum_{r=1}^{R} \frac{\sum\limits_{V \in \mathcal{M}_r} \left[\frac{e^{\sum\limits_i \lambda_i g_i(V, X_r)}}{\sum\limits_{V'' \in \mathcal{M}_r} e^{\sum\limits_i \lambda_i g_i(V'', X_r)}} \right]^\eta \mathcal{L}(V, W_r)}{\sum\limits_{V' \in \mathcal{M}_r} \left[\frac{e^{\sum\limits_i \lambda_i g_i(V', X_r)}}{\sum\limits_{V'' \in \mathcal{M}_r} e^{\sum\limits_i \lambda_i g_i(V'', X_r)}} \right]^\eta} \tag{B.3}$$

$$= \sum_{r=1}^{R} \frac{\sum\limits_{V \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V, X_r)} \right]^\eta \mathcal{L}(V, W_r)}{\sum\limits_{V' \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V', X_r)} \right]^\eta}. \tag{B.4}$$

The partial derivative of $F(\Lambda)$ with respect to the ith model weight λ_i can be obtained by the straightforward application of the multiplication rule:

$$\frac{\partial F(\Lambda)}{\partial \lambda_i} = \tag{B.5}$$
$$\sum_{r=1}^{R} \frac{\sum\limits_{V \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V,X_r)} \right]^{\eta} \eta \; g_i(V,X_r) \; \mathcal{L}(V,W_r) \sum\limits_{V' \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V',X_r)} \right]^{\eta}}{\left\{ \sum\limits_{V' \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V',X_r)} \right]^{\eta} \right\}^2} -$$
$$\sum_{r=1}^{R} \frac{\sum\limits_{V \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V,X_r)} \right]^{\eta} \mathcal{L}(V,W_r) \sum\limits_{V' \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V',X_r)} \right]^{\eta} \eta \; g_i(V',X_r)}{\left\{ \sum\limits_{V' \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V',X_r)} \right]^{\eta} \right\}^2}$$

For the sake of simplicity, an auxiliary quantity is introduced to denote the log-linear distribution with model weights scaled by η:

$$P_{\Lambda \cdot \eta}(V|X_r) = \frac{\left[e^{\sum\limits_i \lambda_i g_i(V,X_r)} \right]^{\eta}}{\sum\limits_{V'' \in \mathcal{M}_r} \left[e^{\sum\limits_i \lambda_i g_i(V'',X_r)} \right]^{\eta}} = \frac{e^{\sum\limits_i (\eta \lambda_i) g_i(V,X_r)}}{\sum\limits_{V'' \in \mathcal{M}_r} e^{\sum\limits_i (\eta \lambda_i) g_i(V'',X_r)}}. \tag{B.6}$$

Inserting Eq. (B.6) into Eq. (B.5) yields:

$$\frac{\partial F(\Lambda)}{\partial \lambda_i} = \sum_{r=1}^{R} \sum_{V \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V|X_r) \; \eta \; g_i(V,X_r) \; \mathcal{L}(V,W_r) -$$
$$\sum_{r=1}^{R} \sum_{V \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V|X_r) \; \mathcal{L}(V,W_r) \sum_{V' \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V'|X_r) \; \eta \; g_i(V',X_r). \tag{B.7}$$

From simply interchanging V and V' in the second term follows:

$$\frac{\partial F(\Lambda)}{\partial \lambda_i} = \sum_{r=1}^{R} \sum_{V \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V|X_r) \; \eta \; g_i(V,X_r) \; \mathcal{L}(V,W_r) -$$
$$\sum_{r=1}^{R} \sum_{V' \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V'|X_r) \; \mathcal{L}(V',W_r) \sum_{V \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V|X_r) \; \eta \; g_i(V,X_r) \tag{B.8}$$

Finally, factoring out the common terms and defining an auxiliary quantity $\tilde{\mathcal{L}}(V, W_r, \Lambda)$ results in the final form of the partial derivative:

$$\frac{\partial F(\Lambda)}{\partial \lambda_i} = \sum_{r=1}^{R} \sum_{V \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V|X_r)\ \eta\ g_i(V, X_r) \cdot \underbrace{\left(\mathcal{L}(V, W_r) - \sum_{V' \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V'|X_r)\ \mathcal{L}(V', W_r) \right)}_{:=\tilde{\mathcal{L}}(V, W_r, \Lambda)} \tag{B.9}$$

$$= \sum_{r=1}^{R} \sum_{V \in \mathcal{M}_r} P_{\Lambda \cdot \eta}(V|X_r)\ \eta\ g_i(V, X_r)\ \tilde{\mathcal{L}}(V, W_r, \Lambda). \tag{B.10}$$

The quantity $\tilde{\mathcal{L}}(V, W_r, \Lambda)$ can be seen as the number of word error with "zero-mean".

Appendix C

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.

C.1 Mathematical Symbols

$a^{[t]}(n)$	first order derivative of the magnitude spectrum of time frame t
$a_t^{(i)}(n)$	ith order derivative of the magnitude spectrum of time frame t
c_{sl}	weight of a Gauss distribution in a Gaussian Mixture Model where s is the state and l is the index of the Gauss distribution
$\delta(i,j)$	Kronecker delta, equals one for $i = j$, and zero otherwise
η	Discriminative Model Combination: scaling factor of initial model scales
ϵ	machine precision
$\phi_f(s)$	state tying index for state s given acoustic feature f
f	frequency
f_s	sample rate
$g(W,X)$	feature function in a log-linear distribution given a word sequence W and a feature vector sequence X
λ	Linear Discriminant Analysis: eigenvalue; Discriminative Model Combination: model weight
$\mathcal{N}(x\|\mu;\Sigma)$	Gauss distribution with mean vector μ and with covariance matrix Σ
$p(s_t\|s_{t-1};w_1^N)$	first order transition probability given the spoken word sequence w_1^N

$p(w\|h)$	language model probability of a word w given a history h
$p(w_1^N\|x_1^T)$	a-posteriori probability for the spoken word sequence w_1^N given the acoustic observations x_1^T
$p(w_1^N)$	language model probability for sentence w_1^N
$p(x_1^T)$	probability for the acoustic observations x_1^T
$p(x_1^T\|w_1^N)$	acoustic emission probability for the acoustic observations x_1^T given the word sequence w_1^N
$p(x_1^T, s_1^T\|w_1^N)$	joint probability for the acoustic observations x_1^T and sequence of Hidden Markov Model states given the word sequence w_1^N
$s^{[t]}$	spectrum derivative measure of time frame t derived form the first order derivative
$s_t^{(i)}$	spectrum derivative measure of time frame t derived form the ith order derivative
s	states of a Hidden Markov Model
s_1^T	temporal sequence of Hidden Markov Model states, also denoted as path
$s(n)$	discrete speech signal
$\tau,\ t$	points in time or time frame indexes
$v_{AC,CC,HPS,AMD}^{[t]}$	voicing measure in time frame t derived from different similarity functions
$w,\ v$	word indexes
w_1^N	two sequences of spoken words $w_1, \ldots, w_N$ and $v_1, \ldots, v_M$
W	signal analysis: window length; discriminative model combination: sequence of words
x_1^T	sequence of acoustic observation vectors $x_1, \ldots, x_T$
$x^{[t]}(\tau)$	time frame t
$x_f^{[t]})$	time frame t produced by feature extraction algorithm f
$y^{[t]}$	Linear Discriminant Analysis: projected feature vector in time frame t
B	between-class scatter matrix
$\mathcal{D}$	Linear Discriminant Analysis: input size

D	Linear Discriminant Analysis: output size
$D^{[t]}(\tau)$	Average Magnitude Difference of time frame t
$E^{[t]}$	energy of time frame t
$F(\omega)$	Filter bank output.
$\mathcal{F}x(t)$	Fourier transform of the signal $x(t)$.
$F(\Lambda, W, X)$	Minimum Word Error for model weights Λ, word sequence W, and feature vector sequence X
$H(\omega)$	Transfer function of the filter of a filter bank.
$\mathcal{L}(V, W)$	number of word errors between word sequences V and W based on Levenshtein-distance
L	Linear Discriminant Analysis: number of concatenated successive feature vectors is $2L + 1$
$\mathcal{M}$	N-best list
$P^{[t]}(n)$	Harmonic Product Spectrum of time frame t
$Q^{[t]}(\tau)$	Average Quadratic Magnitude Difference of time frame t
$P(W)$	prior probability of word sequence W
$P(W\|X)$	a-posteriori probability of word sequence W given a feature vector sequence X (W and X stand often for a complete corpus)
$P_\Lambda(W\|X)$	a-posteriori probability of word sequence W given a feature vector sequence X using the log-linear distribution with model weights Λ (W and X stand often for a complete corpus)
$P(X\|W)$	acoustic emission probability for feature vector sequence X given the word sequence W (W and X stand often for a complete corpus)
$P_f(X_f\|W)$	acoustic emission probability for feature vector sequence X produced by feature extraction algorithm f given the word sequence W (W and X stand often for a complete corpus)
$P(\phi\|W)$	probability of pronunciation variant sequence ϕ given a word sequence W
Φ	pronunciation variant sequence
$R^{[t]}(\tau)$	autocorrelation of time frame t
S	Hidden Markov Model state sequence
$S(\lambda)$	reciprocal condition number of eigenvalue λ

$Difl(x, y)$	reciprocal condition number of eigenvector pair (x, y)
T	number of time frames of an utterance
V	Linear Discriminant Analysis: projector matrix; Discriminative Model Combination: sequence of words
X	Discriminative Model Combination: feature vector sequence
$X(\omega)$	Fourier spectrum
X_f	Discriminative Model Combination: feature vector sequence produced by feature extraction algorithm f
W	Linear Discriminant Analysis: within class scatter matrix; Discriminative Model Combination: sequence of words
$\mathcal{X}$	chordal distanced

C.2 Acronyms

Σ_{LDA}	LDA based combination of features
Σ_{DMC}	DMC based combination of features
am	**a**coustic **m**odel
corr	cross-**corr**elation
cmp	c**omp**onent
del	**del**etion errors
dev	**dev**elopment corpus
em	**em**ission model
eval	**eval**uation corpus
ins	**ins**ertion errors
lm	**l**anguage **m**odel
pr	**pr**onunciation model
sub	**sub**stitution errors
rnd	r**and**om
tr	**tr**ansition model
tr_w	**tr**ansition model within **w**ords
wrd	**wrd** penalty
AC	**A**uto**c**orrelation
AMD	**A**verage **M**agnitude **D**ifference
CART	**C**lassification **A**nd **R**egression **T**ree
CC	**C**ross-**C**orrelation
DMC	**D**iscriminative **M**odel **C**ombination
EM	**E**xpectation-**M**aximization
EPPS	British English **E**uropian **P**arliament **P**lanery **S**ession corpus
FFT	**F**ast **F**ourier **T**ransform
GMM	**G**aussian **M**ixture **M**odel

HMM	**H**idden **M**arkov **M**odel
HPS	**H**armonic **P**roduct **S**pectrum
LDA	**L**inear **D**iscriminant **A**nalysis
LM	**L**anguage **M**odel
LPC	**L**inear **P**rediction **C**oefficients
LVCSR	**L**arge **V**ocabulary **S**peech **R**ecognition
MFCC	**M**el-**F**requency-**C**epstral-**C**oefficients
MF-PLP	Feature using a Mel scale triangular filter bank embedded into the data flow of the PLPfeature
ML	**M**aximum-**L**ikelihood
MLLR	**M**aximum-**L**ikelihood-**L**inear-**R**egression
MVDR	**M**inimum **V**ariance **D**istortionless **R**esponse
MWE	**M**inimum **W**ord **E**rror
NGD	**N**ode **G**raph **D**ensity
PLP	**P**erceptual **L**inear **P**rediction
PP	Language Model **P**er**P**lexity
ROVER	**R**ecognition **O**utput **V**oting **E**rror **R**eduction
RWTH	**R**heinisch **W**estfälische **T**echnische **H**ochschule
SD	**S**pectrum **D**erivative measure (derived from first order derivative)
SIMD	**S**ingle **I**nstruction **M**ultiple **D**ata
V	**V**oicing measure (autocorrelation based)
VM II	**V**erb**M**obil **II** corpus
VTLN	**V**ocal **T**ract **L**ength **N**ormalization
WER	**W**ord **E**rror **R**ate

Bibliography

[Acero 1990] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, September 1990.

[Alleva et al. 1996] P. Alleva, X. D. Huang, and M. Y. Hwang, "Improvements on the pronunciation prefix tree search organization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1996*, Atlanta, GA, USA, May 1996, vol. 1, pp. 133–136.

[Anderson et al. 1999] E. Anderson, Z. Bai, J. Bishop, J. Demmel, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK User's Guide*, SIAM, Philadelphia, PA, 3rd edition, 1999.

[Atal and Rabiner 1976] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201 – 212, 1976.

[Bahl et al. 1983] L. R. Bahl, F. Jelinek, and 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 1975] J. K. Baker, "Stochastic modeling for automatic speech understanding," in *Speech Recognition*, D. R. Reddy, Ed., pp. 512–542. Academic Press, New York, NY, USA, 1975.

[Bakis 1976] R. Bakis, "Continuous speech word recognition via centisecond acoustic states," in *ASA Meeting 1976*, Washington, DC, USA, April 1976.

[Baum 1972] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," in *Inequalities*, O. Shisha, Ed., 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 1957] R. E. Bellman, "Dynamic programming," Princeton University Press, Princeton, NJ, USA, 1957.

[Beulen et al. 1999] K. Beulen, S. Ortmanns, and C. Elting, "Dynamic programming search techniques for across-word modeling in speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1999*, Phoenix, AZ, March 1999, pp. 609–612.

[Beulen 1999] K. Beulen, *Phonetische Entscheidungsbäume für die automatische Spracherkennung mit großem Vokabular*, Ph.D. thesis, RWTH Aachen, July 1999.

[Beyerlein 1997] P. Beyerlein, "Discriminative model combination," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop 1997*, Santa Barbara, CA, December 1997, pp. 238 – 245.

[Beyerlein 1998] P. Beyerlein, "Discriminative model combination," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1998*, Seattle, WA, May 1998, vol. 1, pp. 481 – 484.

[Beyerlein 2000] P. Beyerlein, *Diskriminative Modellkombination in Spracherkennungssystemen mit grossem Wortschatz*, Ph.D. thesis, RWTH Aachen University, October 2000.

[Bocchieri 1993] E. Bocchieri, "Vector quantization for the efficient computation of continuous density likelihoods," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1993*, Minneapolis, MN, April 1993, pp. 692–695.

[Chen et al. 1999] S. S. Chen, E. M. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanevsky, and P. Olsen, "Recent improvements to IBM's speech recognition system for automatic transcription of broadcast news," in *DARPA Broadcast News Workshop 1999*, Herndon, VA, February 1999, pp. 89–93.

[Cook and Robinson 1996] G. Cook and T. Robinson, "Boosting the performance connectionist large vocabulary speech recognition," in *Proc. Int. Conf. on Spoken Language Processing 1996*, Philadelphia, PA, October 1996, vol. 3, pp. 1305 – 1308.

[Davis and Mermelstein 1980] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357 – 366, August 1980.

[Demmel and Kagström 1993a] J. Demmel and B. Kagström, "The generalized schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. part i: Theory and algorithms.," *ACM Transactions on Mathematical Software*, vol. 19(2), pp. 160 – 174, 1993.

[Demmel and Kagström 1993b] J. Demmel and B. Kagström, "The generalized schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. part ii: Theory and algorithms.," *ACM Transactions on Mathematical Software*, vol. 19(2), pp. 175 – 201, 1993.

[Dempster et al. 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[Dharanipragada and Rao 2001] S. Dharanipragada and B.D. Rao, "Mvdr based feature extraction for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2001*, Salt Lake City, UT, May 2001, vol. 1, pp. 309 – 312.

[Doddington et al. 2000] G. R. Doddington, M. A. Przybocki, A. F. Martin, and 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.

[Duda et al. 2001] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2001.

[Eide 2001] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Proc. European Conf. on Speech Communication and Technology 2001*, Aalborg, Denmark, September 2001, pp. 1613 – 1616.

[Ellis and Bilmes 2000] D. Ellis and J. Bilmes, "Using mutual information to design feature combinations," in *Proc. Int. Conf. on Spoken Language Processing 2000*, Beijing, China, October 2000, vol. 3, pp. 79 – 82.

[Evermann et al. 2004] G. Evermann, H.Y. Chan, M.J.F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P.C. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2004*, Montreal, Canada, May 2004, vol. 1, pp. 249–252.

[Fiscus 1997] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop 1997*, Santa Barbara, CA, December 1997, pp. 347 – 354.

[Fisher 1936] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 179-188, 1936.

[Fritsch 1997] J. Fritsch, "ACID/HNN: A framework for hierarchical connectionist acoustic modeling," in *IEEE Workshop on Automatic Speech Recognition and Understanding 1997*, S. Furui, B.-H. Juang, and W. Chou, Eds., Santa Barbara, CA, USA, December 1997, pp. 164–171.

[Generet et al. 1995] M. Generet, H. Ney, and F. Wessel, "Extensions to absolute discounting for language modeling," in *Proc. European Conf. on Speech Communication and Technology 1995*, Madrid, Spain, September 1995, vol. 2, pp. 1245–1248.

[Golub and Loan 1996] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, 3rd edition, 1996.

[Graciarena et al. 2004] M. Graciarena, H. Franco, J. Zheng, D. Vergyri, and A. Stolcke, "Voicing feature integration in sri's decipher lvcsr system," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2004*, Montreal, Canada, May 2004, vol. 1, pp. 921 – 924.

[Gu and Rose 2001] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2001*, Salt Lake City, UT, May 2001, pp. 125 – 128.

[Häb-Umbach and Loog 1999] R. Häb-Umbach and M. Loog, "An investigation of cepstral parameterisations for large vocabulary speech recognition," in *Proc. European Conf. on Speech Communication and Technology 1999*, Budapest, Hungary, September 1999, vol. 3, pp. 1323 – 1326.

[Häb-Umbach and Ney 1992] R. Häb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1992*, San Francisco, CA, March 1992, vol. 1, pp. 13 – 16.

[Häb-Umbach and Ney 1994] R. Häb-Umbach and H. Ney, "Improvements in beam search for 10000-word continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 353–356, April 1994.

[Hegde et al. 2004] R. M. Hegde, H. A. Murthy, and V. Gadde, "Continuous speech recognition using joint features derived from the modified group delay function and mfcc," in *Proc. Int. Conf. on Spoken Language Processing 2004*, Jeju Island, Korea, October 2004, vol. 2, pp. 905 – 908.

[Hegde et al. 2005] R. M. Hegde, H. A. Murthy, and V. Gadde, "Speech processing using joint features derived from the modified group delay function," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2005*, Philadelphia, PA, March 2005, vol. 1, pp. 541 – 544.

[Heigold et al. 2005] G. Heigold, W. Macherey, R. Schlüter, and H. Ney, "Minimum exact word error training," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop 2005*, San Juan, Puerto Rico, Nov. – Dec. 2005.

[Hermansky 1990] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738 – 1752, June 1990.

[Holmes et al. 1997] J. N. Holmes, W. J. Holmes, and P. N. Garner, "Using formant frequencies in speech recognition," in *Proc. European Conf. on Speech Communication and Technology 1997*, Rhodes, Greece, September 1997, vol. 4, pp. 2083 – 2086.

[Hon and Lee 1991] H. W. Hon and K. F. Lee, "Recent progress in robust vocabulary-independent speech recognition," in *DARPA Speech and Natural Language Processing Workshop 1991*, Pacific Grove, February 1991, pp. 258–263.

[Huang and Jack 1989] X. D. Huang and M. A. Jack, "Semi-continuous hidden markov models for speech signals," *Computer Speech and Language*, vol. 3, no. 3, pp. 329–252, 1989.

[Janin et al. 1999] A. Janin, D. Ellis, and N. Morgan, "Multi-stream speech recognition: Ready for prime time?," in *Proc. European Conf. on Speech Communication and Technology 1999*, Budapest, Hungary, September 1999, vol. 2, pp. 591 – 594.

[Jelinek 1969] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM Journal of Research and Development*, vol. 13, pp. 675–685, November 1969.

[Jelinek 1976] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 10, pp. 532–556, April 1976.

[Kanthak et al. 2000] S. Kanthak, K. Schütz, and H. Ney, "Using simd instructions for fast likelihood calculation in lvcsr," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2000*, Istanbul, Turkey, June 2000, pp. 1531–1534.

[Katz et al. 2002] M. Katz, H-G. Meier, H. Dolfing, and D. Klakow, "Robustness of linear discriminant analysis in automatic speech recognition," in *Proc. Int. Conf. on Pattern Recognition 2002*, Québec, Canada, August 2002, vol. 3, pp. 30371 – 30374.

[Katz 1987] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Speech and Audio Processing*, vol. 35, pp. 400–401, March 1987.

[Kirchhoff 1998] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. Int. Conf. on Spoken Language Processing 1998*, Sidney, Australia, December 1998, pp. 891 – 894.

[Kocharov et al. 2005] D. Kocharov, A. Zolnay, R. Schlüter, and H. Ney, "Articulatory motivated acoustic features for speech recognition," in *Proc. European Conf. on Speech Communication and Technology 2005*, Lisboa, Portugal, September 2005, vol. 2, pp. 1101 – 1104.

[Kumar and Andreou 1998] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283 – 297, December 1998.

[Lee and Rose 1996] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1996*, Atlanta, GA, May 1996, vol. 1, pp. 353 – 356.

[Levinson et al. 1983] S. E. Levinson, L. R. Rabiner, and 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.

[Lööf et al. 2006] J. Lööf, S. Umesh, and H. Ney, "Vtln warping factor estimation using accumulation of sufficient statistics," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2006*, Toulouse, France, May 2006.

[Lowerre 1976] B. Lowerre, *A Comparative Performance Analysis of Speech Understanding Systems*, Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, PA, 1976.

[Macherey et al. 2005] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. European Conf. on Speech Communication and Technology 2005*, Lisbon, Portugal, September 2005, pp. 2133–2136.

[McDonough 1998] J. W. McDonough, "Speaker normalization with all-pass transforms," Tech. Rep. 28, Center for Language Speech Processing, The Johns Hopkins University, Baltimore, MD, September 1998, (http://www.clsp.jhu.edu/people/jmcd/postscript/all-pass.ps).

[Metze and Waibel 2002] F. Metze and A. Waibel, "A flexible stream architecture for asr using articulatory features," in *Proc. Int. Conf. on Spoken Language Processing 2002*, Denver, CO, September 2002, pp. 2133 – 2136.

[Molau 2003] S. Molau, *Normalization in the Acoustic Feature Space for Improved Speech Recognition*, Ph.D. thesis, RWTH Aachen, Aachen, Germany, 2003.

[Murthi and Rao 1997] M.N. Murthi and B.D. Rao, "Minimum variance distortionless response (mvdr) modeling of voiced speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1997*, Munich, Germany, April 1997, vol. 3, pp. 1687 – 1690.

[Murthy and Gadde 2003] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2003*, Hong Kong, China, April 2003, vol. 1, pp. 68 – 71.

[Nene and Nayar 1996] S. A. Nene and S. K. Nayar, "Closest point search in high dimensions," in *IEEE Conference on Computer Vision and Pattern Recognition 1996*, San Francisco, CA, June 1996, pp. 859–865.

[Ney and Aubert 1994] H. Ney and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," in *Proc. Int. Conf. on Spoken Language Processing 1994*, Yokohama, Japan, September 1994, vol. 3, pp. 1355–1358.

[Ney and Oerder 1993] H. Ney and M. Oerder, "Word graphs: An efficient interface between continuous speech recognition and language understanding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1993*, Minneapolis, MN, USA, April 1993, vol. 2, pp. 119–122.

[Ney et al. 1987] H. Ney, D. Mergel, A. Noll, and A. Paeseler, "A data-driven organization of the dynamic programming beam search for continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1987*, Dallas, TX, USA, April 1987, pp. 833–836.

[Ney et al. 1992] H. Ney, R. Häb-Umbach, B.-H. Tran, and M. Oerder, "Improvements in beam search for 10000-word continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1992*, San Francisco, CA, March 1992, vol. 1, pp. 9–12.

[Ney et al. 1994] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in language modeling," *Computer Speech and Language*, vol. 2, no. 8, pp. 1–38, 1994.

[Ney et al. 1997] H. Ney, S. C. Martin, and F. Wessel, "Statistical language modeling using leaving-one-out," in *Corpus Based Methods in Language and Speech Processing*, S. Young and G. Bloothooft, Eds., pp. 1–26. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[Ney 1984] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 32, no. 2, pp. 263–271, April 1984.

[Ney 1990] H. Ney, "Acoustic modeling of phoneme units for continuous speech recognition," in *Signal Processing V: Theories and Applications, Fifth European Signal Processing Conference*, L. Torres, E. Masgrau, and M. A. Lagunas, Eds., pp. 65–72. Elsevier Science Publishers B. V., Barcelona, Spain, 1990.

[Odell et al. 1994] J. J. Odell, V. Valtchev, P. C. Woodland, and S. J. Young, "A one-pass decoder design for large vocabulary recognition," in *ARPA Spoken Language Technology Workshop 1994*, Plainsboro, NJ, March 1994, pp. 405–410.

[Ortmanns and Ney 1995] S. Ortmanns and H. Ney, "An experimental study of the search space for 20000-word speech recognition," in *Proc. European Conf. on Speech Communication and Technology 1995*, Madrid, Spain, September 1995, vol. 2, pp. 901–904.

[Ortmanns et al. 1996a] S. Ortmanns, H. Ney, and A. Eiden, "Language-model look-ahead for large vocabulary speech recognition," in *Proc. Int. Conf. on Spoken Language Processing 1996*, Philadelphia, PA, October 1996, vol. 4, pp. 2095–2098.

[Ortmanns et al. 1996b] S. Ortmanns, H. Ney, A. Eiden, and N. Coenen, "Look-ahead techniques for improved beam search," in *CRIM-FORWISS Workshop 1996*, Montreal, Canada, October 1996, pp. 10–22.

[Ortmanns et al. 1997a] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43–72, January 1997.

[Ortmanns et al. 1997b] S. Ortmanns, H. Ney, and T. Firzlaff, "Fast likelihood computation methods for continuous mixture densities in large vocabulary speech recognition," in *Proc. European Conf. on Speech Communication and Technology 1997*, Rhodes, Greece, September 1997, vol. 1, pp. 139–142.

[Ortmanns 1998] S. Ortmanns, *Effiziente Suchverfahren zur Erkennung kontinuierlich gesprochener Sprache*, Ph.D. thesis, RWTH Aachen, Aachen, Germany, November 1998.

[Padmanabhan 2000] M. Padmanabhan, "Spectral peak tracking and its use in speech recognition," in *Proc. Int. Conf. on Spoken Language Processing 2000*, Beijing, China, October 2000, vol. 1, pp. 604 – 607.

[Paliwal 2003] K.K. Paliwal, "Usefulness of phase in speech processing," in *IPSJ Spoken Language Processing Workshop 2003*, Gifu, Japan, February 2003, pp. 1 – 6.

[Parlett and Chen 1990] B. N. Parlett and H. C. Chen, "Use of indefinite pencils for computing damped natural modes," *Journal of Linear Algebra and Its Applications*, pp. 140:53 – 88, 1990.

[Paul 1991] D. B. Paul, "Algorithms for an optimal A^* search and linearizing the search in the stack decoder," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1991*, Toronto, Canada, May 1991, vol. 1, pp. 693–696.

[Picone et al. 1999] J. Picone, S. Pike, R. Regan, T. Kamm, J. Bridle, L. Deng, Z. Ma, H. Richard, and M. Schuster, "Initial evaluation of hidden dynamic models on conversational speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1999*, Phoenix, AZ, March 1999, vol. 1, pp. 109 – 112.

[Pitz and Ney 2003] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of mfcc," in *Proc. European Conf. on Speech Communication and Technology 2003*, Geneva, Switzerland, September 2003, pp. 1445 – 1448.

[Pitz 2005] M. Pitz, *Investigations on Linear Transformations for Speaker Adaptation and Normalization*, Ph.D. thesis, RWTH Aachen University, 2005.

[Povey and Woodland 2002] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2002*, Orlando, FL, May 2002, vol. 1, pp. 105–108.

[Press et al. 1992] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numeric Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 2nd edition, 1992.

[Rabiner and Juang 1986] L. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 4–16, 1986.

[Rabiner and Schafer 1978] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[Rabiner and Schafer 1979] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Signal Processing Series, Englewood Cliffs, NJ, 1979.

[Rabiner 1989] 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, February 1989.

[Ramasubramansian and Paliwal 1992] V. Ramasubramansian and K. K. Paliwal, "Fast k-dimensional tree algorithms for nearest neighbor search with application to vector quantization encoding," *IEEE Transactions on Speech and Audio Processing*, vol. 40, no. 3, pp. 518–528, March 1992.

[Richards and Bridle 1999] H.B. Richards and J.S. Bridle, "The hdm: A segmental hidden dynamic model of coarticulation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1999*, Phoenix, AZ, March 1999, vol. 1, pp. 357 – 360.

[Sakoe 1979] H. Sakoe, "Two-level DP-matching - a dynamic programming-based pattern matching algorithm for connected word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 27, pp. 588–595, December 1979.

[Saon et al. 2003] G. Saon, G. Zweig, B. Kingsburyand L. Mangu, and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Proc. European Conf. on Speech Communication and Technology 2003*, Geneva, Switzerland, September 2003, pp. 1977–1980.

[Schlüter and Ney 2001] R. Schlüter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2001*, Salt Lake City, UT, May 2001, vol. 1, pp. 133 – 136.

[Schwartz and Austin 1991] R. Schwartz and S. Austin, "A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1991*, Toronto, Canada, May 1991, vol. 1, pp. 701–704.

[Schwartz and Chow 1990] R. Schwartz and Y.-L. Chow, "The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1990*, Albuquerque, NM, April 1990, pp. 81–84.

[Schwartz et al. 2004] 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, and L. Chen, "Speech recognition in multiple languages and domains: The 2003 BBN/LIMSI ears system," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2004*, Montreal, Canada, May 2004, vol. 3, pp. 753–756.

[Shire 2001] M.L. Shire, "Multi-stream asr trained with heterogeneous reverberant environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2001*, Salt Lake City, UT, May 2001, vol. 1, pp. 253 – 256.

[Sixtus 2003] A. Sixtus, *Across-Word Phoneme Models for Large Vocabulary Continuous Speech Recognition*, Ph.D. thesis, RWTH Aachen, January 2003.

[Steinbiss et al. 1993] V. Steinbiss, H. Ney, R. Häb-Umbach, B.H. Tran, U. Essen, R. Kneser, M. Oerder, H.G. Meier, X. Aubert, C. Dugast, and D. Geller, "The philips research system for large-vocabulary continuous-speech recognition," in *Proc. European Conf. on Speech Communication and Technology 1993*, Berlin, Germany, September 1993, pp. 2125–2128.

[Stuttle and Gales 2002] M.N. Stuttle and M.J.F. Gales, "Combining a gaussian mixture model front end with mfcc parameters," in *Proc. Int. Conf. on Spoken Language Processing 2002*, Denver, Co, September 2002, vol. 3, pp. 1565 – 1568.

[Thomson and Chengalvarayan 1998] D.L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition feature," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1998*, Seattle, WA, May 1998, vol. 1, pp. 21 – 24.

[Tolba et al. 2002] H. Tolba, S. A. Selouani, and D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2002*, Orlando, FL, May 2002, vol. 1, pp. 837 – 840.

[Umesh et al. 2005] S. Umesh, A. Zolnay, and H. Ney, "Implementing frequency-warping and vtln through linear transformation of conventional mfcc," in *Proc. European Conf. on Speech Communication and Technology 2005*, Lisboa, Portugal, September 2005, vol. 1, pp. 269 – 272.

[Vintsyuk 1971] T. K. Vintsyuk, "Elementwise recognition of continuous speech composed of words from a specified dictionary," *Kibernetika*, vol. 7, pp. 133–143, March 1971.

[Viterbi 1967] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, 1967.

[Wakita 1977] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel identification," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1977*, April 1977, vol. ASSP-25, pp. 183 – 192.

[Weber et al. 2001] K. Weber, H. Bourlard, and S. Bengio, "Hmm2- extraction of formant features and their use for robust asr," in *Proc. European Conf. on Speech Communication and Technology 2001*, Aalborg, Denmark, September 2001, pp. 607 – 610.

[Wegmann et al. 1996] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1996*, Atlanta, GA, May 1996, vol. 1, pp. 339–341.

[Welling and Ney 1996] L. Welling and H. Ney, "A model for efficient formant estimation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1996*, Atlanta, GA, May 1996, vol. 2, pp. 797 – 800.

[Welling et al. 2002] L. Welling, S. Kanthak, and H. Ney, "Speaker adaptive modeling by vocal tract normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 415 – 426, September 2002.

[Welling 1999] L. Welling, *Merkmalsextraktion in Spracherkennungssystemen für grossen Wortschatz*, Ph.D. thesis, RWTH Aachen University, 1999.

[Wessel et al. 1997] F. Wessel, S. Ortmanns, and H. Ney, "Implementation of word based statistical language models," in *Spoken Queries in European Languages (SQEL) Workshop on Multi-Lingual Information Retrieval Dialogs 1997*, Pilsen, Czech Republic, April 1997, pp. 55–59.

[Wölfel et al. 2003] M. Wölfel, J. McDonough, and A. Waibel, "Minimum variance distortionless response on a warped frequency scale," in *Proc. European Conf. on Speech Communication and Technology 2003*, Geneva, Switzerland, September 2003, pp. 1021–1024.

[Woodland et al. 1997] P.C. Woodland, M.J.F. Gales, D. Pye, and S.J. Young, "Broadcast news transcription using htk," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1997*, Munich, Germany, April 1997, vol. 2, pp. 719 – 722.

[Y. Gao and Xiang 2000] J. Huang Y. Gao, R. Bakis and B. Xiang, "Multistage coarticulatory model combining articulatory, formant and cepstral features," in *Proc. Int. Conf. on Spoken Language Processing 2000*, Beijing, China, October 2000, pp. 25 – 28.

[Young 1992] S. J. Young, "The general use of tying in phoneme based HMM recognizers," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1992*, San Francisco, CA, March 1992, vol. 1, pp. 569–572.

[Zhu and Paliwal 2004] D. Zhu and K.K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2004*, Montreal, Quebec, May 2004, vol. 1, pp. 125 – 128.

[Zolnay et al. 2002] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," in *Proc. Int. Conf. on Spoken Language Processing 2002*, Denver, CO, September 2002, vol. 2, pp. 1065 – 1068.

[Zolnay et al. 2003] A. Zolnay, R. Schlüter, and H. Ney, "Extraction methods of voicing feature for robust speech recognition," in *Proc. European Conf. on Speech Communication and Technology 2003*, Geneva, Switzerland, September 2003, vol. 1, pp. 497 – 500.

[Zolnay et al. 2005] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2005*, Philadelphia, PA, March 2005, vol. 1, pp. 457 – 460.

Lebenslauf - Curriculum Vitae

Persönliche Angaben

Name:	András Zolnay
Anschrift:	Pijnboomstraar 2, 4814 KW Breda, Die Niederlande
Telefon:	+31 76 528 42 53
E-Mail:	andras.zolnay@planet.nl
Nationalität:	ungarisch
Geburtsdatum/-ort:	13. Juni 1975 in Budapest
Familienstand:	verheiratet

Ausbildung

Schule

1981 – 1989	Samueli Tibor, Grundschule, Budapest
1989 – 1993	Kaffka Margit, Gymnasium, Budapest Abschluß: Abitur

Studium

1993 – 1998	Studium der Elektrotechnik an der Technischen Universität Budapest. Diplomarbeit zum Thema "Verwirklichung eines verteilten, plattformunabhängigen Vorgangsteuerungssystems". Abschluß mit "sehr gut"
2000 – 2005	Promotion Informatik am Lehrstuhl für Informatik 6 der RWTH Aachen. Dissertation zum Thema "Acoustic Feature Combination for Speech Recognition". Abschluß mit "sehr gut"

Berufstätigkeit

September 1996 – Juni 1998	Wissenschaftliche Hilfskraft am Lehrstuhl für Automatisierung, Technische Universität Budapest
Juli 1998 – Juli 2000	Softwareentwickler am Philips Speech Processing in Wien
September 2000 – Oktober 2005	Wissenschaftlicher Angestellter am Lehrstuhl für Informatik 6 der RWTH Aachen
Seit März 2006	Forscher an der R&D Abteilung von Océ Technologies

Sprachen

Ungarisch:	Muttersprache
Deutsch:	Fließend in Wort und Schrift
Englisch:	Fließend in Wort und Schrift
Niederländisch:	Fließend in Wort und Schrift