MICHAEL KLEINSCHMIDT

ROBUST SPEECH RECOGNITION BASED ON SPECTRO-TEMPORAL PROCESSING

Bibliotheks- und Informationssystem der Universität Oldenburg
2003

Verlag/Druck/
Vertrieb:

Bibliotheks- und Informationssystem
der Carl von Ossietzky Universität Oldenburg
(BIS) – Verlag –
Postfach 25 41, 26015 Oldenburg
Tel.: 0441/798 2261, Telefax: 0441/798 4040
e-mail: verlag@bis.uni-oldenburg.de
Internet: www.bis.uni-oldenburg.de

ISBN 3-8142-0873-0

Preface

Why are even the most advanced computers not able to understand speech nearly half as well as human beings? Even though the rapidly growing performance of microprocessors has enabled speech technology to exhibit major, revolutionary advancements within the last decades, we still are not able to communicate with a computer as naturally as, e. g., the heroes in old-fashioned science fiction movies like 'Star Trek'. The thesis by Michael Kleinschmidt attempts to give an answer to this open question and shows one fascinating way to improve the performance of automatic speech recognition in noisy acoustical environments - an ubiquitous condition that occurs inevitably when machines and regular working conditions come into play.

One basic answer that Michael provides is that properties of the ear have to be taken into account before any progress in automatic speech recognition can be made. In fact, our normal ear is capable of separating the desired speech sounds from even a very noisy background noise scenario without any major problems - an achievement that is still far out of reach for any modern automatic speech recognition system. Interestingly, only a few very basic processing properties of the auditory system appear be the main cause for this great achievement. Hence, it looks promising to transform some of these basic processing mechanisms of the human ear into appropriate speech preprocessing techniques that serve as a front-end to a speech recognition system - and this is the basic idea behind the current thesis.

At first, Michael Kleinschmidt refines the sigma-pi-cell-technique first introduced by Gramß and Strube, e. g. the technique of comparing the energy value at a certain centre frequency with the time-delayed energy averaged across an appropriate region at a different centre frequency. A carefully selected set of these special feature detectors already proves to be rather successful (see chapters three and four of the current thesis).

However, 'Le mieux est l'ennemi du bien'[1]: Later on in his thesis, Michael comes up with an even better feature set (e.g., a set of Gabor functions) that is much more common in visual physiology and psychophysics than in acoustics. These Gabor features implement the idea of a complex, "second order" spectro-temporal feature extractor by considering combinations of temporal and spectral transitions as the template for "desired" speech elements. This approach is not only equivalent to recent findings from neurophysiology and human psychophysics. It also helps Michael to construct a state-of-the-art speech recognition system that is based on the speech recogniser of the international computer science institute (ICSI) in Berkeley, California. Needless to say that this system competes well with the most elaborated automatic speech recognition systems worldwide (see chapters seven and eight) - but read yourself!

Michael is the 28th Ph.D. candidate from a series of excellent graduate students who finished their Ph.D. work at my lab in Oldenburg - and he is among the fastest ones. In addition, he is a charismatic character who is able to infect everybody else in the lab with his high motivation to discover new things. This special talent to motivate younger colleagues and to instantaneously make new friends in the scientific community was also very beneficial during his half-year-stay at the ICSI in Berkeley where I had the chance to share a bit of the "Californian Experience" with him. But not only the Californian sun inspired his work - please read yourself! You may find out that - besides exhaustive work by a few human brains and many computers - a great deal of enthusiasm and inspiration is involved in this thesis that originates from the desire to solve the riddles in hearing theory imposed by mother nature herself. May all these positive attributes of the current thesis contribute to our common long-term goal: to provide computers with functional ears!

Birger Kollmeier, March 2003

[1] Voltaire: 'The best is the enemy of the good'

Contents

1 General Introduction **9**

1.1 The Auditory Approach to Signal Processing 9

1.2 Front Ends for Automatic Speech Recognition 12

1.3 Spectro-temporal Modulation Detection 14

1.4 SNR Estimation 18

1.5 Structure of this Thesis 19

2 Combining Speech Enhancement and Auditory Feature Extraction for Robust Speech Recognition **21**

2.1 Introduction 23

2.2 Auditory Model 26

2.3 Digit Recognition Experiments with PEMO Front End . . 28

2.3.1 Setup 28

2.3.2 Results 29

2.4 Digit Recognition Experiments with Monaural Speech Enhancement and PEMO Front End 33

2.4.1 Setup 33

2.4.2 Results 34

2.5 Digit Recognition Experiments with Binaural Speech Enhancement and PEMO Front End 38

2.5.1 Setup 38

2.5.2 Results 39

2.6 Direct Comparison of Monaural and Binaural Speech Enhancement Methods 42
2.6.1 Setup 42
2.6.2 Results 43
2.7 Discussion 45
2.8 Outlook 48

3 Sigma-pi Cells as Secondary Features for Isolated Digit Recognition 51
3.1 Introduction and Summary 51
3.2 Recognition System 52
3.2.1 Perceptual Feature Extraction 53
3.2.2 Secondary Feature Extraction 53
3.2.3 Feature Set Optimization 54
3.3 Experiments 55
3.3.1 Optimal Number and Types of Features 55
3.3.2 Robustness 57
3.4 Summary and Outlook 57

4 Sigma-pi Cells as Secondary Features for Phoneme Recognition 61
4.1 Introduction 62
4.2 Description of Sigma-pi Cells 62
4.3 Why Using Sigma-pi Cells ? 63
4.4 Experimental Setup 64
4.4.1 Corpus 64
4.4.2 Feature Extraction 65
4.4.3 Secondary Feature Extraction 66
4.4.4 Fisher Score 67
4.4.5 Classification and Feature Selection 67
4.5 Evaluation 68
4.5.1 Feasibility 68

4.5.2 Parameter Optimization 68
4.5.3 Feature Analysis 69
4.6 Discussion 72

5 Sub-band Signal-to-noise-ratio Estimation Using Auditory Feature Processing 73
5.1 Introduction 74
5.2 Feature Extraction 77
5.2.1 Material 77
5.2.2 Primary Feature Extraction 78
5.2.3 Sigma-pi Cells as Secondary Features 81
5.3 Classifier 84
5.3.1 Feature-finding Neural Network 84
5.3.2 Multi-layer Perceptron 86
5.4 Experiments 87
5.4.1 Setup 87
5.4.2 Results 89
5.4.3 Primary Feature Dependency 92
5.4.4 Secondary Feature Dependency 93
5.4.5 Segment Length Dependency 96
5.4.6 Computational Effort 97
5.4.7 Importance of Temporal Modulation 98
5.5 Discussion 100

6 Methods for Capturing Spectro-temporal Modulations in Automatic Speech Recognition 103
6.1 Introduction 104
6.2 Secondary Features 107
6.2.1 Sigma-pi Cells 107
6.2.2 Fuzzy Logic Units 110
6.2.3 Gabor Receptive Fields 111
6.3 Automatic Speech Recognition Experiments 114

6.3.1 Material 114
6.3.2 Primary Feature Extraction 114
6.3.3 Recognizer 115
6.3.4 Results 117
6.4 Discussion 118

7 Spectro-temporal Gabor Features as a Front End for Automatic Speech Recognition 121
7.1 Introduction 122
7.2 Gabor Filter Functions 124
7.3 Feature Selection 126
7.4 ASR Experiments 132
7.5 Summary 133

8 Improving Word Accuracy with Gabor Feature Extraction 137
8.1 Introduction 138
8.2 Spectro-temporal Feature Extraction 139
8.3 ASR Experiments 141
8.3.1 Setup 141
8.3.2 Feature Selection 143
8.3.3 Results 144
8.4 Conclusion 147

9 Summary and Conclusions 149

Bibliography 155

A Tables and Figures 163

CHAPTER 1

General Introduction

One of the greatest challenges in acoustical signal processing is the reliable transcription of spoken language into written words, usually referred to as automatic speech recognition (ASR). Another is the separation of the one-dimensional input time signal (pressure at the ear drum or microphone) into its components or streams and the association of each stream with its origin. The latter problem is known as auditory scene analysis (ASA; Bregman, 1990) and is in its simplest form just a measure of the speech-to-noise ratio (SNR) of the input signal. Both problems are connected in many ways. In adverse acoustical conditions, for example, ASA can be considered a prerequisite for successful speech recognition. It is commonly observed in both fields that normal hearing human listeners by far outperform any technical approach to date. The work described in this thesis mainly focuses on improving ASR systems by introducing new auditory features. In addition, this auditory approach is also applied to long-term sub-band SNR estimation.

1.1 The Auditory Approach to Signal Processing

The importance and possibilities of ASR have increased in recent years due to the dramatic progress made in miniaturization of electronic devices. Applications of ASR should facilitate human-machine interaction and communication, thereby making the use of machines accessible to everyone including small children, the visually impaired and elderly people. Furthermore, it allows for automatic transcription into written language of e.g. meetings, broadcasts and telecommunication. Although some progress

has been made in the last twenty years of research and development, ASR is not omnipresent in today's world.

The main reason why ASR is still not applicable on a large scale is that state-of-the-art ASR technology does only work, i.e. yields high recognition rates, in very controlled situations. It can be used for dictation systems, for example, because the amount of reverberation and additive noise is usually limited and the classifier has been trained on the individual speaker and recording system beforehand. If the acoustical conditions vary by nature of the application, e.g. voice-directed menus in automated answering systems, the number of words in the inventory has to be reduced to very small numbers (usually 10-100 items).

Techniques which make the ASR system more *robust* against (less affected by) interfering sound sources and reverberation are generally classified according to the part of the processing chain they target (Gong, 1995). *Preprocessing* of the acoustical signal can lead to speech enhancement and yields a time signal with a better SNR. *Feature extraction* methods are located in the front end of the recognition system, while *model adaptation* schemes are applied within the back end or classifier itself. Figure 1.1 sketches the processing stages of ASR systems which are relevant for this thesis.

The comparison of ASR to normal-hearing native listeners clearly shows that there is still a large gap in performance (Lippmann, 1997). Human listeners recognize speech even in very adverse acoustical environments with strong reverberation and interfering sound sources. Speech *understanding* and knowledge about syntax and semantics of spoken language are highly beneficial in solving a speech recognition task, and the latter are already partly incorporated in ASR technology by language models. In addition, a good internal SNR is necessary for correct classification. The internal representation of speech is the product of many processing steps along the auditory pathway of humans and presented to higher stages of the auditory system superior to the primary auditory cortex (A1). Beginning with the frequency decomposition in the cochlea in the inner ear and continuing with e.g. level adaptation and onset/offset enhancement through the brain stem, midbrain and cortex a complex internal image or representation of the acoustical stimulus is created. The human auditory system is assumed to increase the internal SNR by auditory scene analysis. The auditory stream belonging to the source of interest is selected using cues such as spatial location, co-modulation and pitch. The signal processing carried

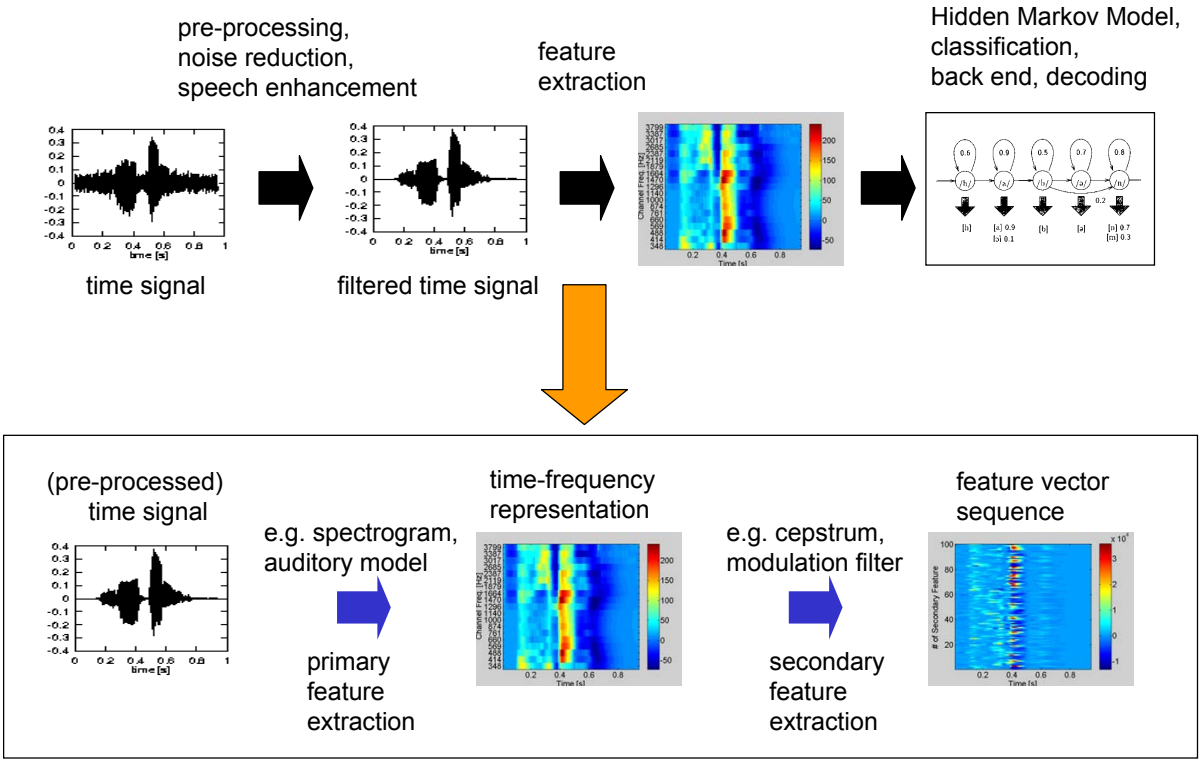


Figure 1.1: Schematic overview of the possible processing steps carried out in an automatic speech recognition system: The original time signal can be subject to speech enhancement/noise reduction in the pre-processing step. Primary feature extraction is then carried out on the resulting filtered time signal, yielding a spectro-temporal representation in form of a primary feature vector sequence. Another part of the front end may be a modulation filtering step resulting in a sequence of secondary feature vectors. Those are then fed into the back end/classifier/recognizer, which normally consists of a Hidden Markov Model (HMM) for Gaussian mixture modeling of likelihoods and a subsequent decoding step.

out by the human auditory system is not yet known in detail. Nevertheless, ongoing research in psychoacoustics and neurophysiology is making progress and some models exist about human auditory processing that leads to a presumably very robust internal representation of speech in A1 and on higher stages.

The approach pursued in this thesis is to learn from the biological blueprint of a very good recognition system, i.e. from models of the human auditory system. After all, we use speech mainly to be understood by other human beings. Therefore, the idea of the auditory approach to technical speech processing is to mimick certain key elements of the human auditory system and to use the psychoacoustical and neurophysiological knowledge to build better algorithms. However, a blind one-to-one copy of biologi-

cal structures is not only computationally impractical. In his plea for the auditory approach in ASR, Hermansky (1998) draws an analogy to the development of flying machines (i.e. airplanes), which do not flap their wings to fly, but rely on a specific shape of their wings. He concludes that "...progress should be made by the *knowledge of the principle guiding a process*, rather than by *copying the appearance of a process.*" This means for ASR that one has to investigate which parts of auditory processing are important for speech perception and how these may be incorporated into the statistical framework of state-of-the-art ASR technology.

1.2 Front Ends for Automatic Speech Recognition

The standard method for feature extraction are mel-cepstral coefficients. Coefficients of a short-term Fourier transform of the time-windowed speech signal are combined into several frequency channels according to the logarithmic mel-scale. The results are log-compressed in amplitude/energy and a cosine transform is applied across frequency channels. After that only the first 10-15 coefficients are kept, corresponding to the lowest spectral modulations. The spectral analysis, as well as the logarithmic compression in frequency and amplitude are in fact a very basic model of human auditory processing. For the biological system, the spectral decomposition is carried out in the cochlea and tonotopy is preserved all the way up the auditory pathway. The cepstrum and a related technique called perceptual linear prediction (PLP; Hermansky, 1990) explicitly target formant frequencies. The formants are maxima of the spectral envelope of speech, produced by resonances of the vocal tract and charateristic for individual vowels.

These standard front ends thereby only represent the spectrum within short analysis frames and tend to neglect very important dynamic patterns in the speech signal. This deficiency has been partly overcome by adding dynamic features, i.e., approximations to temporal derivatives in the form of delta and delta-delta features to the set (Furui, 1986).

There is a long history of attempts to utilize computational models based on psychoacoustical and neurophysiological data as front ends for ASR (e.g. Ghitza 1988; Seneff 1988). In recognition experiments, however, these auditory-based front ends often yield only small or no improvements compared to standard front ends, or require high computational costs

(Jankowski et al., 1995). Many rather technical approaches turned out to be related to psychoacoustical models. Examples are the modulation spectrogram (Kingsbury et al., 1998) or the RelAtive SpecTrAl (RASTA) log-domain band pass filtering technique. The latter was originally hand-designed to reduce convolutive channel effects (Hermansky and Morgan, 1994). Later filter coefficients were data-derived and turned out to be very similar to the hand designed ones (Avendano et al., 1996). The long time constant of the bandpass filter has another striking effect: RASTA processing models the psychoacoustical effect of forward masking (Hermansky and Pavel, 1998). A major effect of auditory processing is a band pass filtering of the envelope of the input signal with a best frequency of about 4 Hz and a pass band between around 2 and 16 Hz. This results in onset and offset enhancement and introduces a certain sluggishness in the processing. This sluggishness seems to be beneficial, probably because it matches the physical limitations of the speech production system, which are reflected in the modulation characteristic of speech. The modulation filtering approach not only enhances the overall SNR for additive noise but also reduces convolutive channel effects when applied in the logarithmic frequency domain.

A new candidate for auditory feature extraction for ASR is the model of auditory perception (PEMO) after Dau et al. (1996a). It was originally developed to predict human performance in typical psychoacoustical spectral and temporal masking experiments. The temporal properties are due to five non-linear adaptation loops with time constants between 5 and 500ms which perform near logarithmic compression for stationary signals and linear processing for rapid changes. PEMO was later extended by incorporating a modulation filterbank (Dau et al., 1997a). Its applications include different tasks in the field of speech processing such as the prediction of speech quality and intelligibility (Hansen and Kollmeier 1997; Holube and Kollmeier 1996). The usefulness of PEMO as a front end of ASR was first investigated by Tchorz and Kollmeier (1999a). They reported increased recognition performance in isolated word recognition experiments by replacing the standard mel-cepstrum front end with PEMO and analyzed the importance of individual model components for this application. Further studies showed that the benefit of the PEMO front end holds especially in combination with a locally-recurrent neural network (LRNN) classifier (Tchorz et al., 1997; Kasper et al., 1997; Tchorz et al., 1999) or when applying a linear transformation before the Hidden Markov Model back end (Kasper and Reininger, 1999). In this thesis, the PEMO front

end is further improved by combining it with preceding speech enhancement methods. It is also used for ASR in combination with a secondary feature extraction stage that explicitly targets spectro-temporal envelope fluctuations.

1.3 Spectro-temporal Modulation Detection

In the beginning of the 20th century, Fletcher and colleagues examined speech intelligibility of human listeners for the nascient telecommunication industry. They found log sub-band classification error probability to be additive for nonsense syllable recognition tasks. This suggests independent processing in a number of articulatory bands without recombination until a very late processing stage. Their work resulted in the definition of the articulation index, a model of human speech perception (Fletcher, 1953). When Allen (1994) published a review of that work, the development of a new class of feature extraction methods for ASR was inspired. Instead of calculating cepstra or other features over the whole frequency range, features were now extracted in individual sub-bands. The most extreme example of the new type of purely temporal features are the TRAPs (Sharma, 1999; Hermansky and Sharma, 1998) which utilize multi-layer perceptrons (MLP) to classify current phonemes in each single critical band based on a temporal context of up to 1s. Another approach is multi-band processing (Bourlard et al., 1996a), for which localized cepstral features are calculated in broader sub-bands to reduce the effect of band-limited noise on the overall performance.

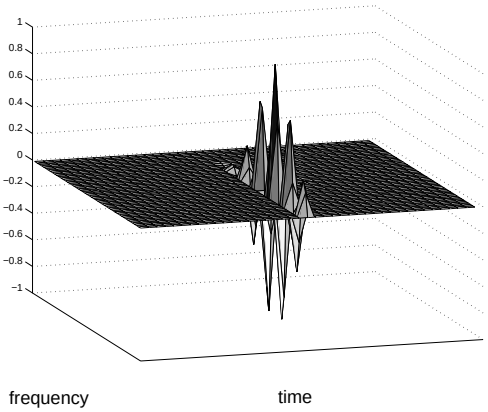
All these feature extraction methods described above apply either spectral or temporal processing at a time. Nevertheless, speech and many other natural sound sources exhibit distinct spectro-temporal amplitude modulations. While the temporal modulations are mainly due to the syllabic and phonetic structure of speech, resulting in a bandpass characteristic with a peak around 4Hz (Kanedera et al., 1999; Chi et al., 1999), spectral modulations describe the harmonic and formant structure of speech. The latter are not at all stationary over time. Coarticulation and prosody result in variations of fundamental and formant frequencies even within a single phoneme. This raises the question whether there is relevant information in amplitude variations oblique to the spectral and temporal axes and

how these diagonal structures may be utilized to improve the performance of automatic classifiers. Figure 1.2 sketches examples of purely spectral, purely temporal and spectro-temporal feature extraction.

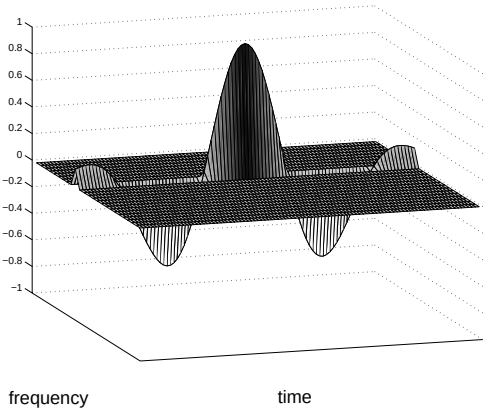
In addition, recent speech intelligibility experiments showed that the combination of two distant narrow spectral channels or slits leads to a gain in intelligibility which is greater than predicted by the articulation index (Greenberg et al., 1998; Warren and Bashford, 1999; Müsch and Buus, 2001). This synergistic effect of distant spectral channels leads to doubts concerning the log error additivity inherent to the articulation index. In his experiments, Fletcher implicitly assumed continuity along the spectrum by only working with high pass and low pass filters. The new data suggests some integration of information across frequency bands. This is supported by a number of physiological experiments on different mammal species which have revealed the spectro-temporal receptive fields (STRF) of neurons in the primary auditory cortex. Individual neurons are sensitive to specific spectro-temporal patterns in the incoming sound signal. The results were obtained using reverse correlation techniques with complex spectro-temporal stimuli such as checkerboard noise (deCharms et al., 1998) or moving ripples (Schreiner and Calhoun, 1994; Kowalski et al., 1996). The STRF often clearly exceed one critical band in frequency, have multiple peaks and also show tuning to temporal modulation (Schreiner et al., 2000). In many cases, the neurons are sensitive to the direction of spectro-temporal patterns (e.g. upward or downward moving ripples), which indicates combined spectro-temporal processing rather than consecutive stages of spectral and temporal filtering (Depireux et al., 2001). These findings fit well to psychoacoustical evidence of early auditory features (Kaernbach, 2000), yielding patterns that are distributed in time and frequency and in some cases comprised of several unconnected parts.

These STRF and early auditory features can be approximated, although somewhat simplified, by sigma-pi cells (the product of two windows in the spectro-temporal plane) or alternatively by two-dimensional Gabor functions, which are localized sinusoids known from receptive fields of neurons in the visual cortex (De-Valois and De-Valois, 1990). Both, the sigma-pi cells and the Gabor function have been investigated as secondary feature sets in this thesis. They are called 'secondary features' for being calculated from a spectro-temporal representation (primary feature matrix) of the input signal. Therefore, sigma-pi cells and Gabor filter functions both target two-dimensional envelope fluctuations. The aim of this thesis is to develop an improved feature extraction method for ASR and signal classi-

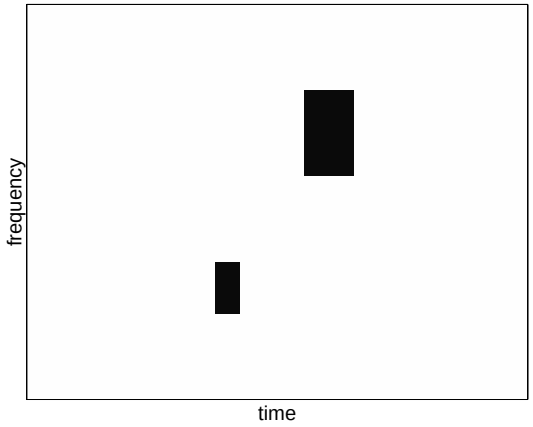
a) spectral feature



b) temporal feature



c) sigma-pi cell



d) Gabor filter

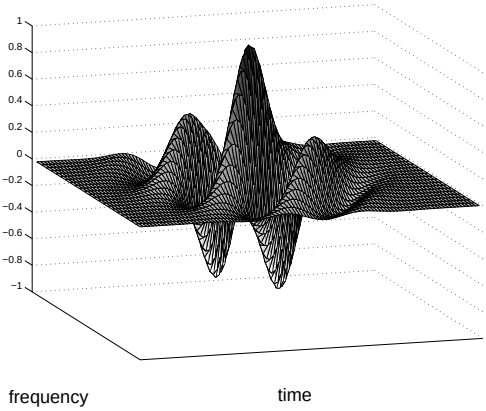


Figure 1.2: Spectral, temporal and spectro-temporal processing for feature extraction is sketched on a two-dimensional spectro-temporal representation of primary feature vectors: a) purely spectral processing is applied e.g. for the mel-cepstrum or PLP features, while b) temporal processing refers to the TRAPs approach. Combined spectro-temporal integration of information is carried out with c) sigma-pi cells and d) Gabor filter functions. a) and b) are plotted as special cases of the Gabor feature class that resemble a localized cepstrum and a TRAPs-like temporal filtering, respectively.

fication. Motivated by the above psychoacoustical and neurophysiological findings, methods of spectro-temporal modulation detection and filtering are designed and evaluated on several ASR tasks. One of them is also applied to long-term SNR estimation in individual frequency bands.

There are few other approaches of spectro-temporal processing for ASR. Nadeu et al. (2001) applied filtering in the two-dimensional Fourier domain of a spectrogram representation but then returned to purely spectral fea-

tures for classification. Many approaches to use artifical neural networks for ASR classify spectral features using temporal context of the order of 10 to 100ms. Depending on the system, this is part of the back end as in the connectionist approach (Bourlard and Morgan, 1998) or part of the feature extraction as in the Tandem system (Hermansky et al., 2000). Still, the temporal context is relatively short and diagonal structures are not explicitly derived. Kajarekar et al. (2001) used linear discriminant analysis (LDA) to obtain data-derived features in the spectro-temporal domain with 1s of context. For phone targets, the first linear discriminants are purely spectral and span the whole frequency range. Some purely temporal features play a minor role but no diagonal features are found. Somervuo (2002) compared LDA in a similar task to non-linear discriminant analysis and independent component analysis (ICA). The resulting features always performed better in phone recognition experiments than the baseline cepstral features with first derivatives. The feature set obtained via ICA on phone targets for 100ms of context contained some more complex spectro-temporal patterns. In another approach proposed by Lin et al. (2000) a genetic algorithm was applied to find sets of optimal features based on a two-dimensional cepstral representation. Automatic phonetic transcription based on articulatory feature detection in spectrogram representation of 1s extent is pursued by Chang et al. (2000, 2001a,b).

Sigma-pi cells were first proposed as secondary features for ASR by Gramß and Strube (1990) and later investigated as part of the Feature-finding Neural Network (FFNN; Gramß, 1992). The FFNN is composed of a linear classifier and an automated feature selection scheme. Gramß (1991) developed a number of learning rules for feature optimization and very efficient algorithms for the selection process. Retrospectively, those algorithms can be labeled 'wrapper methods' for feature selection using the terminology of John et al. (1994), because the importance of individual features is derived on a classification task and in context with the other features in the set. The large number of possible parameter combinations is one of the problems of using spectro-temporal representations as features. This issue may be solved implicitly by automatic learning in neural networks with a spectrogram input and a long time window of e.g. 1s. However, this is computationally expensive and prone to overfitting as it requires large amounts of (labeled) training data, which is often unavailable. By putting further constraints on the spectro-temporal patterns, the number of free parameters can be decreased by orders of magnitude. This is carried out when Gabor functions or sigma-pi cells are used as features.

This approach narrows the search to a certain sub-set and thereby some important features might be ignored, but the constraints are based on neurophysiological and psychoacoustical data. In this thesis, the sigma-pi and Gabor methods of secondary feature extraction are investigated. Feature selection is mainly performed within the FFNN framework using a linear classifier. Analysis of the resulting feature sets is carried out to assess the importance of explicitly targeting diagonal structures/combined spectro-temporal modulation in the front end. In the application the optimized feature set is often combined with a more sophisticated back end, which is better capable of dealing with the temporal variance in speech and estimating continuous target values.

1.4 SNR Estimation

As stated above, SNR estimation can be regarded as a simple form of auditory scene analysis. A good estimate of a long-term (> 500ms) and global (whole frequency range) SNR provides an estimate of the acoustical situation. This is important for ASR and other applications such as telecommunication and digital hearing aids. Many complex algorithms for speech processing are optimized for specific acoustic environments and might produce artifacts in situations that do not meet the specific assumptions. This is the case for speech enhancement techniques, compression algorithms in hearing aids, and robust feature extraction methods and model compensation techniques in ASR. Some of these applications perform band-wise processing and therefore require an estimate of the sub-band SNR, i.e. the SNR in each frequency band. SNR estimates with a higher temporal resolution of up to one frame are needed to steer a noise reduction scheme (such as Wiener filtering or spectral subtraction) directly. The high temporal resolution is also necessary in ASR to use the missing data approach (time and frequency localized missing values, Cooke et al., 2001) or the frame dropping technique that omits whole frames with very low SNR (Adami et al., 2002).

Dupont and Ris (1999, 2001) compared different methods for SNR estimation, most of them based on the amplitude statistics (Martin, 1993; Hirsch, 1993; Hirsch and Ehrlicher, 1995; Bourlard et al., 1996a) over a longer period of time (up to 1s). Tchorz and Kollmeier (1999b, 2001) introduced a method for broad band and sub-band SNR estimation that is based solely on amplitude modulation spectrograms (AMS) segments, which are 32ms

long. The resulting representation clearly exhibits characteristics related to the harmonic and formant structure of speech (if present). Tchorz et al. (2001) applied this AMS-based SNR estimator to noise reduction for ASR applications. In this thesis, sub-band long-term SNR estimation is carried out by means of the sigma-pi cell approach to ASR, which targets low frequency spectro-temporal envelope fluctuations only.

1.5 Structure of this Thesis

This thesis consists of seven main chapters that document the development and application of auditory feature extraction to problems of signal classification. After evaluating the combination of noise reduction techniques and an auditory model for feature extraction in ASR, sigma-pi cells are introduced as a secondary feature extraction method and applied to ASR and SNR estimation. The sigma-pi cell approach is further developed into refined Gabor filters for spectro-temporal modulation detection. The complexity of the investigated tasks and back ends increases during the course of this thesis, from German digit recognition with simple recognition systems to sub-word classification and multi-lingual digit string recognition with state-of-the-art classifiers.

2 Based on the encouraging results by Tchorz and Kollmeier (1999a), the experiments in **Chapter 2** are carried out to investigate how speech enhancement techniques in the pre-processing stage further increase the performance of the PEMO/LRNN system in isolated digit recognition experiments. Monaural noise reduction, as proposed by Ephraim and Malah (1984), is compared to a binaural filter and de-reverberation algorithm after Wittkop et al. (1997) in noisy and reverberant environments.

3 In **Chaper 3** the auditory model is extended by using sigma-pi cells as secondary features for spectro-temporal modulation detection. While in earlier studies with sigma-pi cells by Gramß and Strube (1990) contrasted Bark spectrograms were used as primary input features, the combination of PEMO as primary input features and sigma-pi cells/FFNN is experimentally analyzed here for isolated digit recognition in adverse noise conditions.

4 **Chapter 4** deals with the usefulness of sigma-pi features for shorter segments of speech, single phonemes. The feature sets obtained by the automatic feature selection scheme are analyzed.

5 The PEMO/sigma-pi front end for ASR and the FFNN feature selection framework described in Chapter 3 may also be utilized to estimate the long-term sub-band SNR of a sound signal based on low-frequency envelope fluctuations. This is documented in **Chapter 5**.

6 In **Chapter 6**, the isolated digit recognition experiments are extended to different variations of sigma-pi cells with more than two windows and more general stochastic combinations of windows. In addition, the new Gabor filter function is proposed, investigated and compared to the window-based sigma-pi approaches.

7 The Gabor approach is further developed in **Chapter 7** by optimizing the feature sets on phoneme, diphone and digit target labels. The resulting sets of Gabor filters are statistically analyzed and incorporated into a MLP/HMM Tandem system for automatic recognition of digit strings.

8 Finally, in **Chapter 8**, the Gabor-based Tandem system is improved by noise reduction techniques and combined with and compared to other feature streams obtained from state-of-the-art front ends. The systems are evaluated within the Aurora framework for small corpora, a multilingual digit string recognition task.

Detailed, complete result tables and some extra figures are given in **Appendix A**.

This thesis is organized in self-contained chapters that are suitable for publication as independent journal articles. This composition should convince the reader that auditory-based processing strategies have a great potential for ASR and other application areas in speech communication. Especially the use of spectro-temporal features will be shown to improve the performance of ASR systems in adverse acoustical conditions.

CHAPTER 2

Combining Speech Enhancement and Auditory Feature Extraction for Robust Speech Recognition [a]

Contents

2.1 Introduction 23
2.2 Auditory Model 26
2.3 Digit Recognition Experiments with PEMO Front End 28
2.4 Digit Recognition Experiments with Monaural Speech Enhancement and PEMO Front End 33
2.5 Digit Recognition Experiments with Binaural Speech Enhancement and PEMO Front End 38
2.6 Direct Comparison of Monaural and Binaural Speech Enhancement Methods 42
2.7 Discussion 45
2.8 Outlook 48

Abstract

A major deficiency in state-of-the-art automatic speech recognition (ASR) systems is the lack of robustness in additive and convolutional noise. The model of auditory perception (PEMO), developed by Dau et al. (1996a) for psychoacoustical purposes, partly overcomes these difficulties when used as a front end for automatic speech recognition. To further improve the performance of this auditory-based recognition system in background noise, different speech enhancement methods were examined, which have been

[a] A slightly modified version of this chapter was published in *Speech Communication* (34) **1–2**, pp.75–91 (2001) by Michael Kleinschmidt, Jürgen Tchorz and Birger Kollmeier.

evaluated in earlier studies as components of digital hearing aids. Monaural noise reduction, as proposed by Ephraim and Malah (1984), was compared to a binaural filter and dereverberation algorithm after Wittkop et al. (1997). Both noise reduction algorithms yield improvements in recognition performance equivalent to up to 10dB SNR in non-reverberant conditions for all types of noise, while the performance in clean speech is not significantly affected. Even in real-world reverberant conditions the speech enhancement schemes lead to improvements in recognition performance comparable to an SNR gain of up to 5dB. This effect exceeds the expectations as earlier studies found no increase in speech intelligibility for hearing-impaired human subjects.

Zusammenfassung

Die mangelnde Robustheit moderner Systeme zur automatischen Spracherkennung gegenüber additiven und konvolutiven Störungen ist eines der drängensten Probleme aktueller Forschung. Das Perzeptionsmodell nach Dau et al. (1996a), welches ursprünglich für psychoakustische Anwendungen konzipiert wurde, kann als auditorische Vorverarbeitung zu einer robusteren Erkennungsleistung beitragen. Um die Klassifikationsleistung dieses gehörbasierten Erkennungssystems weiter zu erhöhen, wurden verschiedene Methoden zu Störgeräuschunterdrückung untersucht, welche in der Vergangenheit als Komponenten digitaler Hörgeräte evaluiert wurden. Verglichen wurde das monaurale Verfahren zur Störgeräuschreduktion nach Ephraim und Malah (1984) mit dem binauralen Filter und Enthallungsalgorithmus nach Wittkop et al. (1997). In reflexionsarmer Umgebung bewirkten beide Algorithmen eine Erhöhung der Erkennungsleistung, entsprechend einer Verbesserung des Signal-Rausch-Abstands um bis zu 10dB für alle untersuchten Störgeräusche, während die Ergebnisse in Ruhe nicht beeinträchtigt wurden. Selbst in realer, verhallter Umgebung erreichten die Störunterdrückungsverfahren Verbesserungen der Erkennungsleistung vergleichbar einem um bis zu 5dB günstigeren SNR. Diese Ergebnisse übertreffen die Erwartungen, da in früheren Untersuchungen für schwerhörige Versuchspersonen mit digitalen Hörgeräten keine Erhöhung der Sprachverständlichkeit gefunden werden konnte.

2.1 Introduction

A major problem of most automatic speech recognition (ASR) systems is their unsatisfactory robustness in noise. Several researchers proposed front ends which simulate different processing stages of the auditory system to overcome this problem, as human 'feature extraction' leads to very robust speech understanding in noise (Ghitza 1988; Seneff 1988). In recognition experiments, however, these auditory-based front ends often yield only small or no improvements compared to standard front ends, or require high computational costs (Jankowski et al., 1995). A further approach of auditory feature extraction is investigated here. It is based on a model of the auditory periphery (PEMO) which was originally developed by Dau et al. (1996a) to predict human performance in typical psychoacoustical masking experiments, but was also applied to different tasks in the field of speech processing (Hansen and Kollmeier 1997; Holube and Kollmeier 1996). It has been shown that using PEMO as a front end for automatic speech recognition systems results in additional robustness compared to standard mel-frequency cepstral coefficient (MFCC) front ends (Tchorz and Kollmeier, 1999b), especially when applying locally-recurrent neural networks (LRNN) as classifiers for isolated word recognition tasks (Tchorz et al., 1997).

Another method to overcome the lack of robustness observed for state-of-the-art ASR systems is to enhance the incoming time signal before feature extraction. A number of single-channel noise reduction algorithms have been examined as pre-processing steps for ASR (recent work e.g. Mine et al. 1996; Fischer and Stahl 1999; Gelin and Junqua 1999; Hermus et al. 1999; Vizinho et al. 1999). Since MFCC-based recognition systems are prone to degradation of performance on clean speech when combined with speech enhancement schemes (Kermorvant and Morris 1999; Wilmers and Strube 1999), the robustness of a given ASR system against distortions introduced by noise reduction is of major concern. Multi-Channel approaches towards speech enhancement for ASR often consist of physically large microphone arrays (Kiyohara et al. 1997; Omologo et al. 1997; Bitzer et al. 1999). A special type of multi-channel processing is the *binaural* approach, i.e., a two channel approach that assumes two microphones positioned in the 'ears' of a dummy head[b] or probe microphones close to the ears of a real person. Since this approach allows the simulation of the binaural signal

[b]Or alternatively to the left and right of a roughly head-sized and head-shaped object.

processing and noise reduction usually present in normal-hearing listeners, it has attracted much attention in the area of auditory modeling (Durlach 1972; Colburn 1996; Blauert 1997; Zerbs 1999), noise reduction for hearing impaired listeners (Kollmeier et al. 1993; Peissig and Kollmeier 1997; Wittkop et al. 1997) and ASR (Bodden and Anderson, 1995; Francis and Anderson, 1997; Kleinschmidt et al., 1998).

This paper describes the benefit the combined PEMO/LRNN system may gain by employing several methods of speech enhancement, both monaural and binaural. Single-channel noise reduction algorithms such as the minimum mean square error short-term spectral amplitude estimator (Ephraim and Malah, 1984) rely on temporal windows in which speech is absent to reestimate the quasi-stationary noise spectrum. Two channel algorithms in general require more technical effort, but have the possibility of directional filtering and dereverberation by exploiting the differences in phase and level between the two signals recorded at the left and the right hand side of a head-like object. Both types of noise reduction algorithms have been applied to the task of increasing speech intelligibility for hearing-impaired listeners (Marzinzik and Kollmeier 1999; Wittkop et al. 1999) - but only showed a limited benefit. Although the signal-to-noise ratio (SNR) is improved by monaural and binaural speech enhancement in certain laboratory experiments, the algorithms are of limited use for humans in realistic acoustic environments. In most situations, speech intelligibility was not significantly improved, or even degraded. Some benefits could be observed in terms of 'ease of listening' and listening fatigue. One important reason for these limited benefits is that the algorithms have been used at comparatively unfavorable SNRs, where normal-hearing listeners still understand speech quite well, while impaired listeners have tremendous difficulties. ASR systems, on the other hand, have even more problems with additive noise than hearing-impaired listeners, since their performance declines at much more favorable SNRs, where noise reduction schemes yield a higher benefit. It is therefore worthwhile to combine the noise reduction strategies primarily developed for digital hearing aids with robust ASR systems to achieve an even better performance in noise. In this paper, monaural and binaural algorithms therefore have been tested for a number of types of noise and signal-to-noise ratios, keeping a constant PEMO front end and LRNN recognizer. The aim of this study was, on the one hand, to obtain a more robust speech recognition system and, on the other hand, an objective evaluation of the speech enhancement schemes.

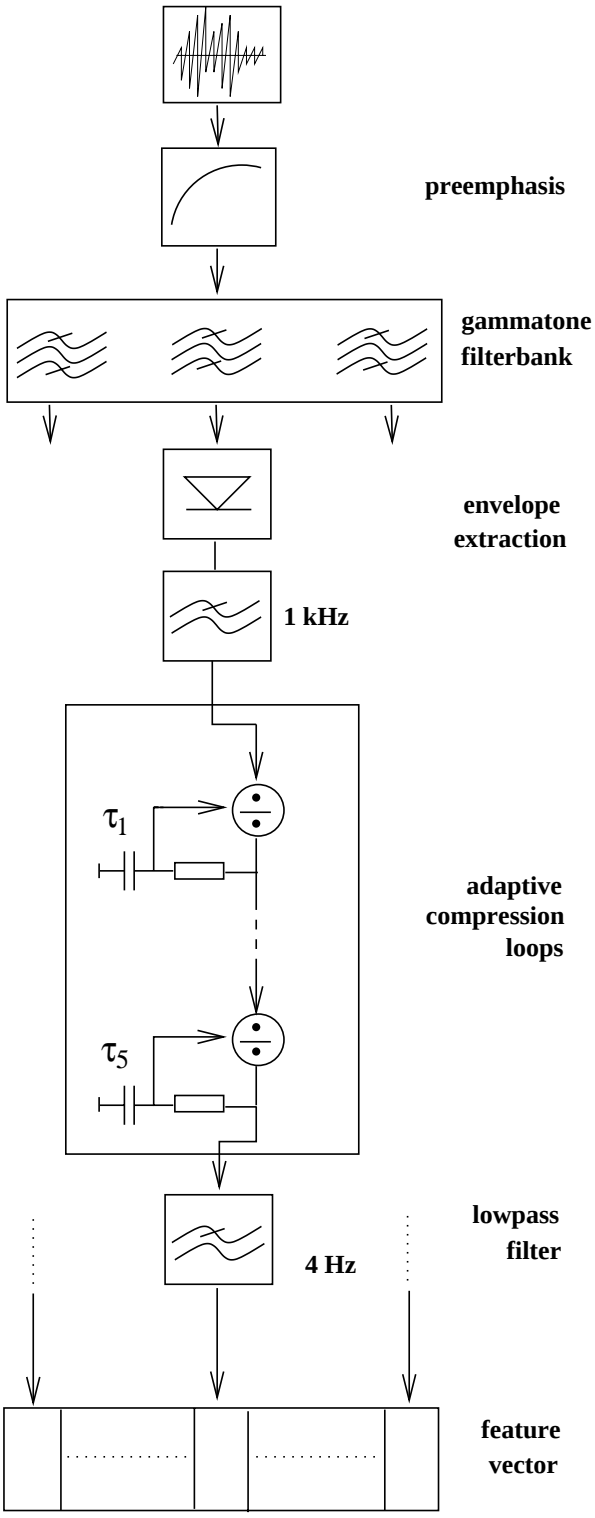


Figure 2.1: Processing stages of the auditory model (PEMO). The initial time signal is high pass filtered (pre-emphasis, 7kHz Butterworth) and then spectrally decomposed by a bank of gammatone band pass filters (one ERB spacing and width). In each frequency channel the envelope is then extracted by half-way rectification and low pass filtering (1kHz). After that, a series of five non-linear adaptation loops performs compression and constrasting of onsets and offsets. Each loop consists of a divider which uses its low pass filtered output as the denominator (low pass time constants: 5, 50, 129, 253 and 500Hz). Finally a modulation low ass (4Hz) is applied and averaging yields a vector of output values every 10ms.

2.2 Auditory Model

The model of auditory perception (PEMO) by Dau et al. (1996a) was designed as a model of the 'effective' signal processing that takes place in the auditory periphery transforming the acoustic signal into its 'internal representation'. It quantitatively accounts for a number of psychoacoustical experiments carried out with human subjects (Dau et al. 1996b, 1997b), e.g. spectral and forward masking, temporal integration and modulation perception. In addition, this model has been successfully applied to the task of objective speech quality measurement (Hansen and Kollmeier, 1997), speech intelligibility prediction in noise (Wesselkamp, 1994) and for hearing-impaired listeners (Holube and Kollmeier 1996; Derleth 1999).

In Figure 2.1 the processing stages of the auditory model are shown. The first processing step is a pre-emphasis of the input signal with a first-order high pass filter. This flattens the typical spectral tilt of speech signals and reflects the transfer function of the outer ear. The preemphasized signal is then filtered by a gammatone filterbank (Patterson et al., 1987) using nineteen frequency channels equally spaced on the ERB scale with center frequencies ranging from 300 to 4000Hz. The impulse responses of the gammatone filterbank are similar to the impulse responses of the auditory system found in physiological measurements. After gammatone filtering, the signal in each frequency channel is halfwave-rectified and first order low pass filtered with a cutoff frequency of 1kHz for envelope extraction, which reflects the limiting phase-locking for auditory nerve fibers above 1kHz. Amplitude compression is performed in a subsequent processing step. In contrast to conventional bank-of-filters front ends, the amplitude compression of the auditory model is not static (e.g., instantaneously logarithmic) but adaptive, which is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops. Each of these loops consists of a divider and an RC low pass filter with an individual time constant ranging from 5 to 500ms. Changes in the input signal like onsets and offsets are emphasized, whereas steady-state portions are compressed. Thus, the dynamical structure of the input signal is taken into account over a relatively long period of time. Short term adaptation including enhancement of changes and temporal integration is simulated and allows a quantitative prediction of important temporal effects in auditory perception.

The last processing step of the auditory model is a first order low pass filter with a cutoff frequency of 4Hz. It attenuates fast envelope fluctuations of

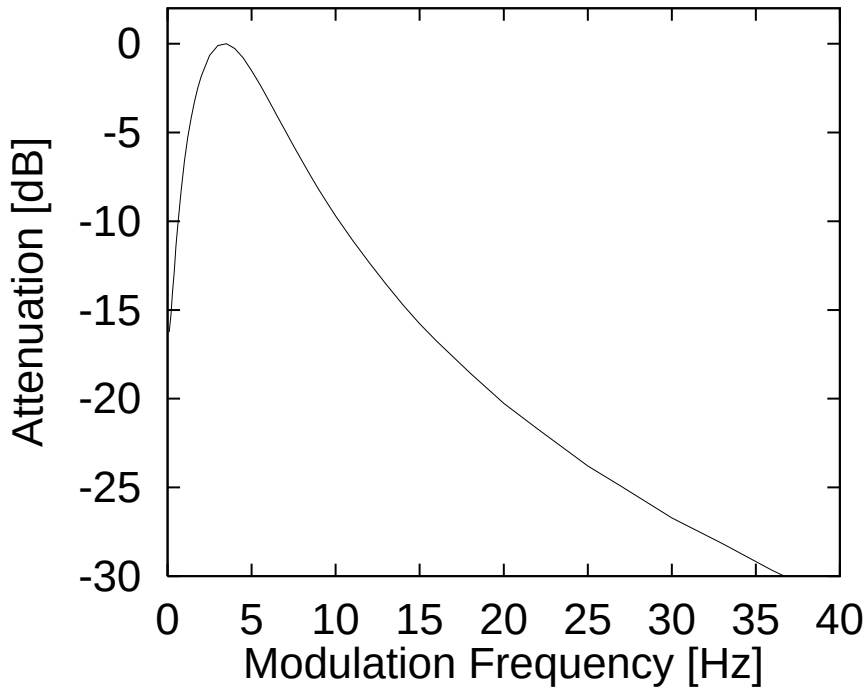


Figure 2.2: Modulation transfer function of PEMO. Due to the nonlinear nature of the adaptation loops the transfer function is signal-dependent. The values plotted here are calculated using an amplitude modulated sinusoidal carrier at 1kHz. Note that the band pass characteristic is due to the combination of modulation low pass and the adaptation loops, which compress stationary signals. Data is taken from Tchorz and Kollmeier (1999a).

the signal in each frequency channel. Suppression of very slow envelope fluctuations by the adaptation loops and attenuation of fast fluctuations by the low pass filter results in a band pass characteristic of the amplitude modulation transfer function of the auditory model with a maximum at about 4Hz (see Figure 2.2). This corresponds well to the average modulation spectrum of speech, which also has its maximum at around 4Hz. An extension of the model (not used here) replaces the final low pass filter by a bank of modulation band pass filters (Dau et al., 1997a). The output of the auditory model is downsampled to a rate of 100 feature vectors per second to serve as input to the recognizer.

2.3 Digit Recognition Experiments with PEMO Front End

In this section results from isolated word recognition experiments are introduced making use of PEMO auditory feature extraction *without* further speech enhancement. While the robustness of the PEMO front end had been documented elsewhere (Tchorz et al., 1997; Kasper et al., 1997; Tchorz and Kollmeier, 1999a; Kasper and Reininger, 1999), the results of this section are intended to serve as a baseline for the following experiments with noise reduction algorithms.

2.3.1 Setup

A number of speaker-*in*dependent, isolated digit recognition experiments in different types of additive noise were carried out to evaluate the robustness of the auditory-based representation of speech quantitatively. The speech material for training the word models and scoring was taken from the ZIFKOM database of Deutsche Telekom AG. Each German digit was spoken once by 200 different speakers (100 females, 100 male). The speech material was equally divided into two parts for training and testing, each consisting of 1000 utterances by 50 male and 50 female speakers. Training of the word models was always performed on clean digits only. Testing was performed on clean and on noisy digits. Two types of noise were added to the utterances with signal-to-noise ratios between 15 and -10dB: a) noise which was generated from a random superposition of phonetically balanced single words from a male speaker (Sotscheck noise, see Kollmeier et al., 1988), and b) unmodulated speech shaped noise (CCITT G.227), with a spectrum similar to the long-term spectrum of speech.

As a control front end, mel-frequency cepstral coefficients (MFCC) were examined, which are widely used in common ASR systems. The FFT-based coefficients were calculated from Hamming-windowed, preemphasized 32ms segments of the input signal with a frame period of 10ms. In our experiments, each Mel cepstrum feature vector contained twenty-six features (twelve coefficients, log energy, and the respective first temporal derivatives).

Two different recognizers were taken for training and testing: 1.) a standard continuous-density HMM recognizer with five Gaussian mixtures per

state, diagonal covariance matrices and six emitting states per word model, and 2.) a locally-recurrent neural network (LRNN) with three layers of neurons (95 or 130 input, 225 hidden, and ten output neurons). Hidden layer neurons have recurrent connections to their twenty-four nearest neighbors. The input matrix consisted of five times the PEMO output vector with nineteen elements, glued together in order to allow the network to memorize the whole time sequence of input matrices. In the case of using the MFCC front end the input layer consisted of five times twenty-six input neurons. For training, the Backpropagation-Through-Time algorithm was applied in 200 iterations (see Kasper et al., 1995, for a detailed description). In total, four different combinations of front ends and recognizers were compared to each other: MFCC/CHMM, MFCC/LRNN, PEMO/CHMM, and PEMO/LRNN.

2.3.2 Results

The speaker-independent digit recognition rates in clean speech and in additive noise obtained with the different combinations of front ends and recognizers are shown in Figure 2.3. The results for CCITT speech shaped noise and for Sotscheck noise are shown on top and bottom, respectively. The recognition rates in per cent are plotted as a function of the signal-to-noise ratio in dB. In clean speech, all combinations yield similar recognition rates (see also Table 2.1 on page 35).

In additive noise the performance diverges. With a HMM recognizer, both front ends yield comparable results. PEMO works slightly better in Sotscheck noise, and about as good as MFCC features in CCITT noise. With the neural network as classifier, however, the choice of the front end is essential for the recognition rates. Cepstral coefficients yield only poor results in noise with the LRNN recognizer, as already reported in earlier studies by Kasper et al. (1997). When combined with the LRNN recognizer PEMO features provide a useful improvement in robustness when compared with the other combinations tested.

Tchorz et al. (1997) found that the distinct peaks in the representation of the speech signals are the most relevant information for the LRNN recognizer. A recognition rate above 90 % is maintained even if the 80 % lowest feature values are set to zero. HMM recognition, on the other hand, shows degraded performance in that experiment. As the threshold for manipulating the features increases, the recognition rate drops rapidly. It seems

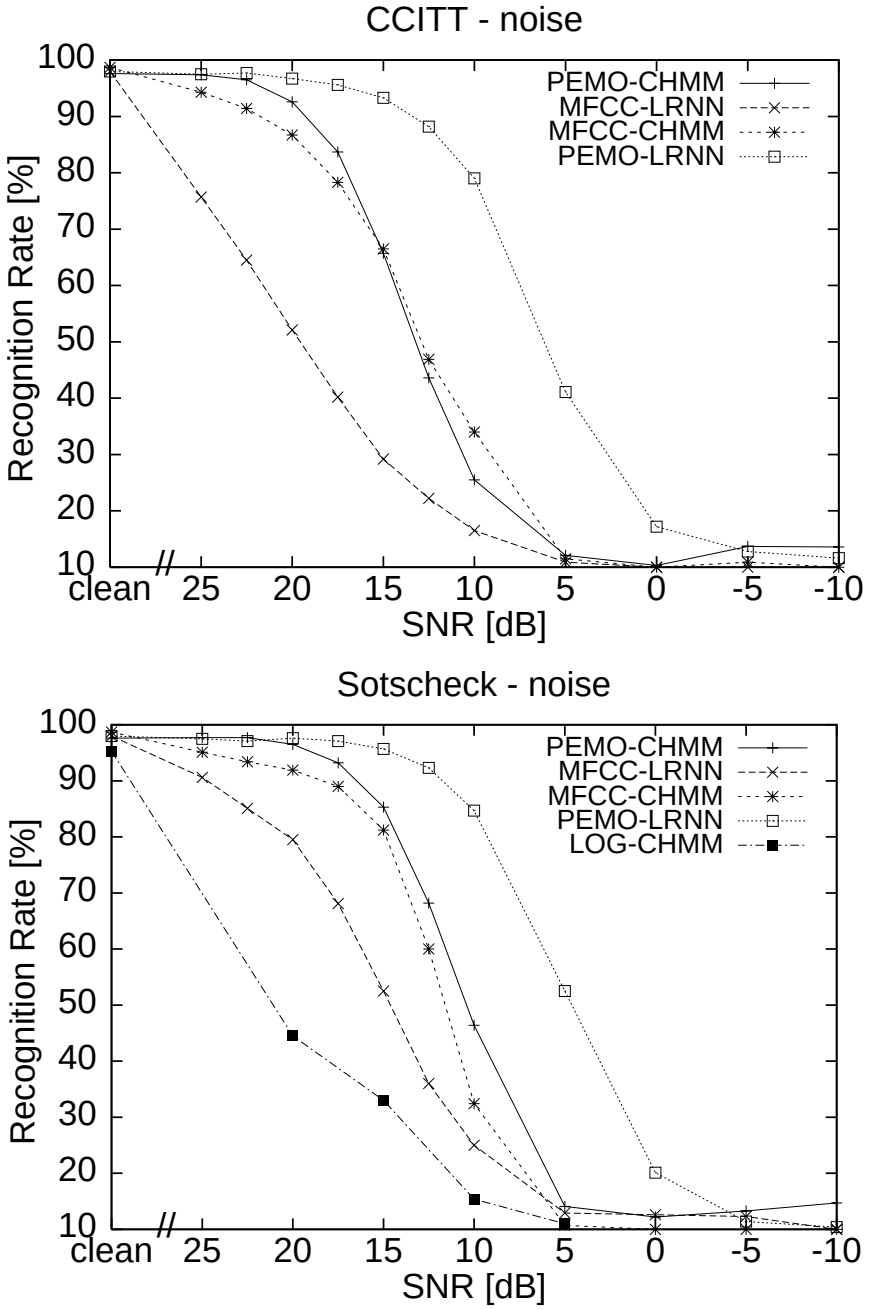


Figure 2.3: Speaker-independent, isolated digit recognition rates in CCITT noise (top) and Sotscheck noise (bottom) as function of SNR for different combinations of front ends (MFCC and PEMO) and recognizers (HMM and LRNN). The data points for condition LOG-CHMM are taken from Tchorz and Kollmeier (1999a) and will be discussed in Section 2.7.

as if HMM recognition exploits all information encoded in the features, including the low values between distinct peaks. These are the parts in the representation which are more distorted in background noise, as can

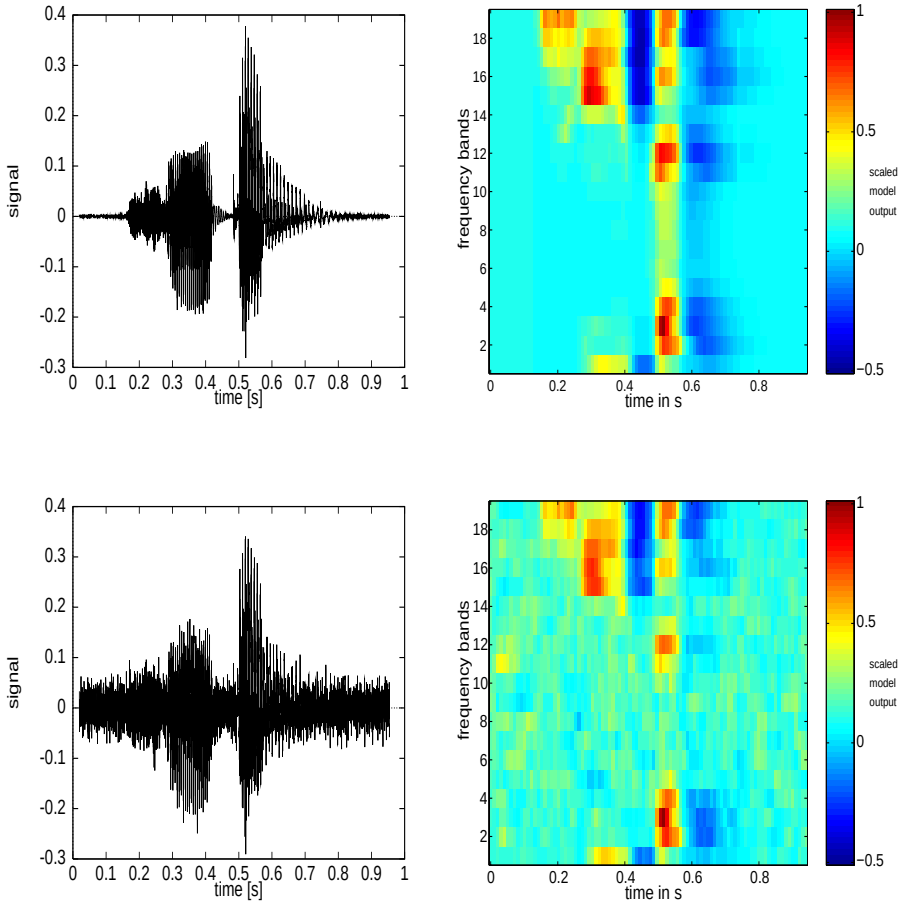


Figure 2.4: Examples for PEMO processing of speech. Top: time signal of an utterance of the German digit sieben (left) and representation after processing (right). Bottom: same utterance mixed with CCITT noise at 5dB SNR and representation after processing.

be seen from Figure 2.4, where PEMO processing is demonstrated without and with the presence of background noise.

While the LRNN seems to benefit from the sparse representation of PEMO features, this might by a problem for HMM recognizers. Also the non-diagonal elements of the covariance matrices of PEMO features are not negligible small. As reported by Kasper and Reininger (1999) the performance of PEMO combined with CHMM recognizers can be further improved by applying a cepstrum-like transformation to the features (thereby obtaining so called PEMO-CEP features). The resulting recognition scores are comparable to the performance of PEMO/LRNN.

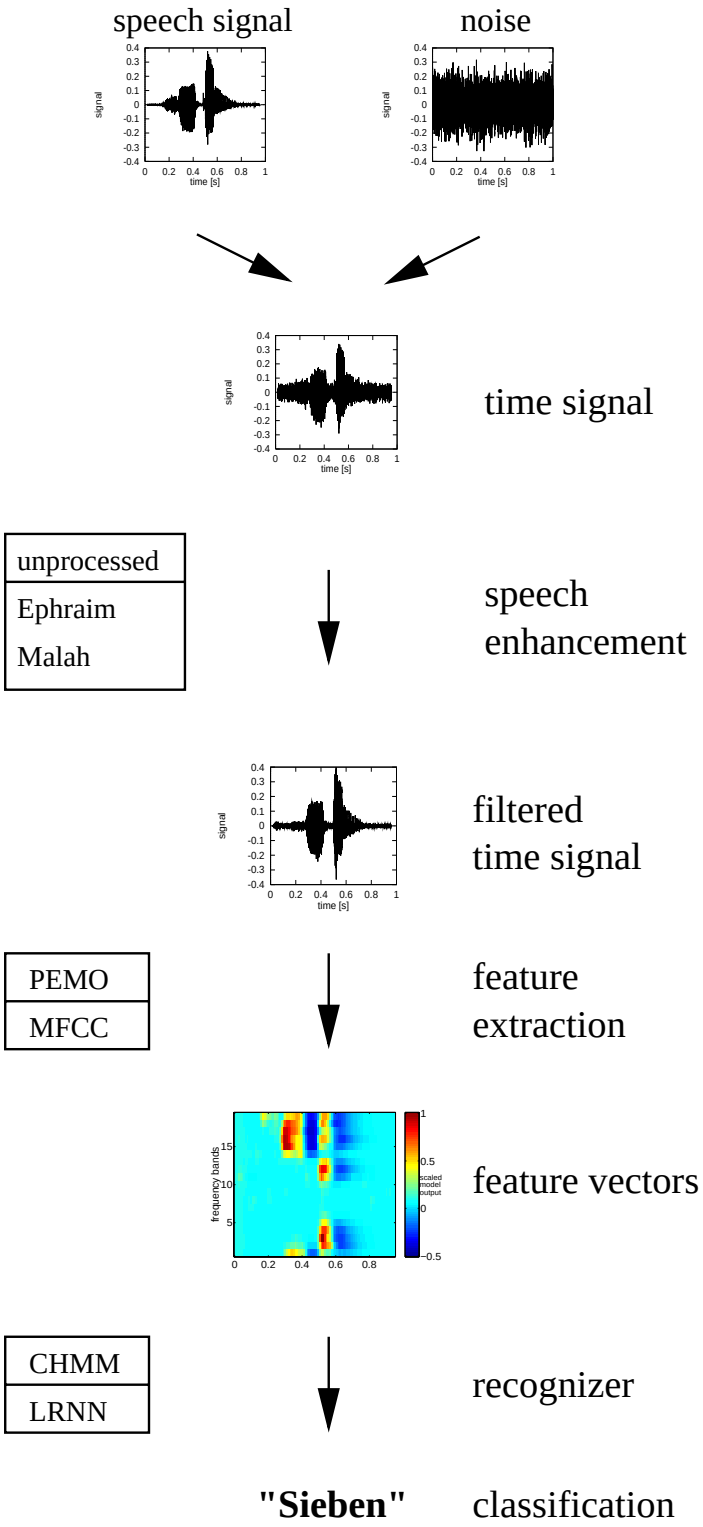


Figure 2.5: Setup of isolated digit recognition experiment. Noise was added to the speech signal, which was then pre-processed via speech enhancement before feature extraction and classification.

2.4 Digit Recognition Experiments with Monaural Speech Enhancement and PEMO Front End

In order to further increase the robustness of the recognition system a single-channel speech enhancement method was added to the experimental setup.

2.4.1 Setup

The minimum mean square error short-term spectral amplitude estimator as proposed by Ephraim and Malah (1984) applies a statistically derived optimal gain to the spectral components. The gain is calculated using estimates of a-posteriori and a-priori SNR (the so-called 'decision directed approach'). This algorithm leads to an audible reduction of additive background noise without distorting the speech signal or producing 'musical tone' artifacts (Cappé, 1994). As with most single-channel noise reduction algorithms, an estimate of the noise spectrum is required. The estimate has to be updated if the additive noise is only quasi-stationary for certain time intervals. Marzinzik and Kollmeier (1999) developed a combination of the Ephraim-Malah scheme and an algorithm to automatically update the noise estimate for use in digital hearing aids. Another study focused on the usefulness of a number of variants of the Ephraim-Malah algorithm regarding its application in robust speech recognition (Kleinschmidt et al., 1999) and showed that the original filter performed better in the given setup than slightly different variants (Ephraim and Malah, 1985), that account for the uncertainty of signal presence or use logarithmic spectral amplitude values.

The experimental setup resembles the one described in Section 2.3 for the previous experiments and is shown in Figure 2.5. The disturbed time signal is filtered by the Ephraim-Malah speech enhancement algorithm as described in Marzinzik and Kollmeier (1999) before feature extraction. In addition to Sotscheck and CCITT noise, construction site noise (from Siemens, 1992) and white Gaussian noise have been used in this comparison. The first 50ms of each signal file were regarded as noise and therefore supplied an initial estimate of the noise spectrum. As isolated digits were used, automatic noise updating did not have to be applied necessarily, but was nevertheless included, since this is indispensable for any application in

realistic environments. As above the training was carried out on one half of the dataset. The training data was left clean and unprocessed.

2.4.2 Results

The speaker-independent digit recognition rates in clean speech and in additive CCITT noise obtained with the different combinations of front ends and recognizers using monaural speech enhancement are shown in Figure 2.6. Again, the recognition rates in per cent are plotted as a function of the signal-to-noise ratio in dB.

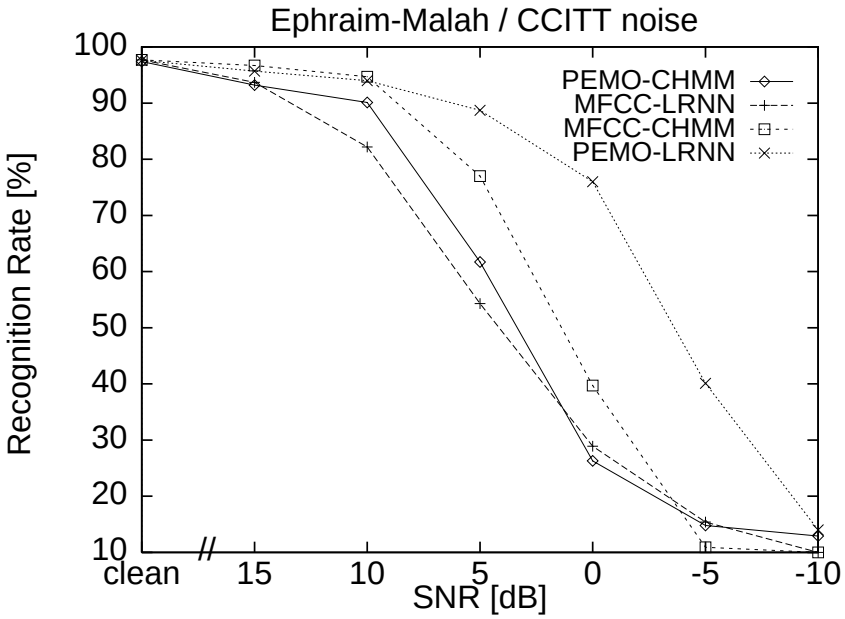


Figure 2.6: Speaker-independent, isolated digit recognition rates in CCITT noise as function of SNR applying Ephraim-Malah speech enhancement for different combinations of front ends (MFCC and PEMO) and recognizers (HMM and LRNN).

The performance on clean test data shows no significant degradation for all combinations of front ends and recognizers, except for MFCC/CHMM where the recognition rate drops by one per cent absolute (see Table 2.1). This result is another hint for a rather 'gentle' noise reduction by the Ephraim-Malah algorithm. Earlier studies (Kleinschmidt et al., 1999; Wilmers and Strube, 1999) have shown the adverse effect of other noise reduction schemes on clean speech error rates, especially for MFCC-based systems. As an overall result the robustness of PEMO-based recognition systems was found to be higher than for MFCC front ends, not only against additive noise, but also the artifacts of speech enhancement processing.

Table 2.1: Speaker-independent isolated word recognition rate in % on clean test data for different combinations of front end and classification tool with Ephraim-Malah speech enhancement and no processing.

	without noise reduction	with Ephraim-Malah
PEMO - CHMM	97.6	97.4
PEMO - LRNN	98.0	97.6
MFCC - LRNN	98.0	97.7
MFCC - CHMM	98.7	97.7

By comparing Figure 2.6 with Figure 2.3 it becomes obvious that all combinations of front ends and recognizers show improved robustness in additive CCITT noise when applying the Ephraim-Malah monaural speech enhancement to the disturbed time signal before feature extraction. PEMO/LRNN is still the most robust with an effective gain of 8dB (at 90 % level) compared to no speech enhancement (cf. Figure 2.3), or a gain of 50 % absolute recognition rate at 5dB SNR level. The following experiments are restricted to the PEMO/LRNN recognition system as this combination appears to be the most promising one.

The speaker-independent digit recognition rates of the PEMO/LRNN combination in different types of additive noise are given in Figure 2.7. The results for unprocessed and Ephraim-Malah filtered signals are located on top and on bottom, respectively.

The robustness against noise of the PEMO/LRNN recognition system significantly depends on the type of background noise added. At most SNR levels white noise seems to have the least effect on recognition performance compared to construction site noise or Sotscheck noise. Adding speech shaped CCITT noise results in the lowest recognition rates. Both CCITT and Sotscheck noise have a smooth spectrum which is similar to the long-term spectrum of speech. In contrast to CCITT, Sotscheck and in particular construction site noise are more modulated types of noise, the latter with high spectral energies at very low and very high frequencies. The results indicate that spectral distribution is a more important factor for the disturbance of speech recognition performance than modulation, at least when moderate modulation depths are compared.

It can be clearly seen in Figure 2.7 that major improvements in robustness are obtained not only for CCITT noise but also for all other types of noise. This effect is less true for construction site noise, with its rather non-

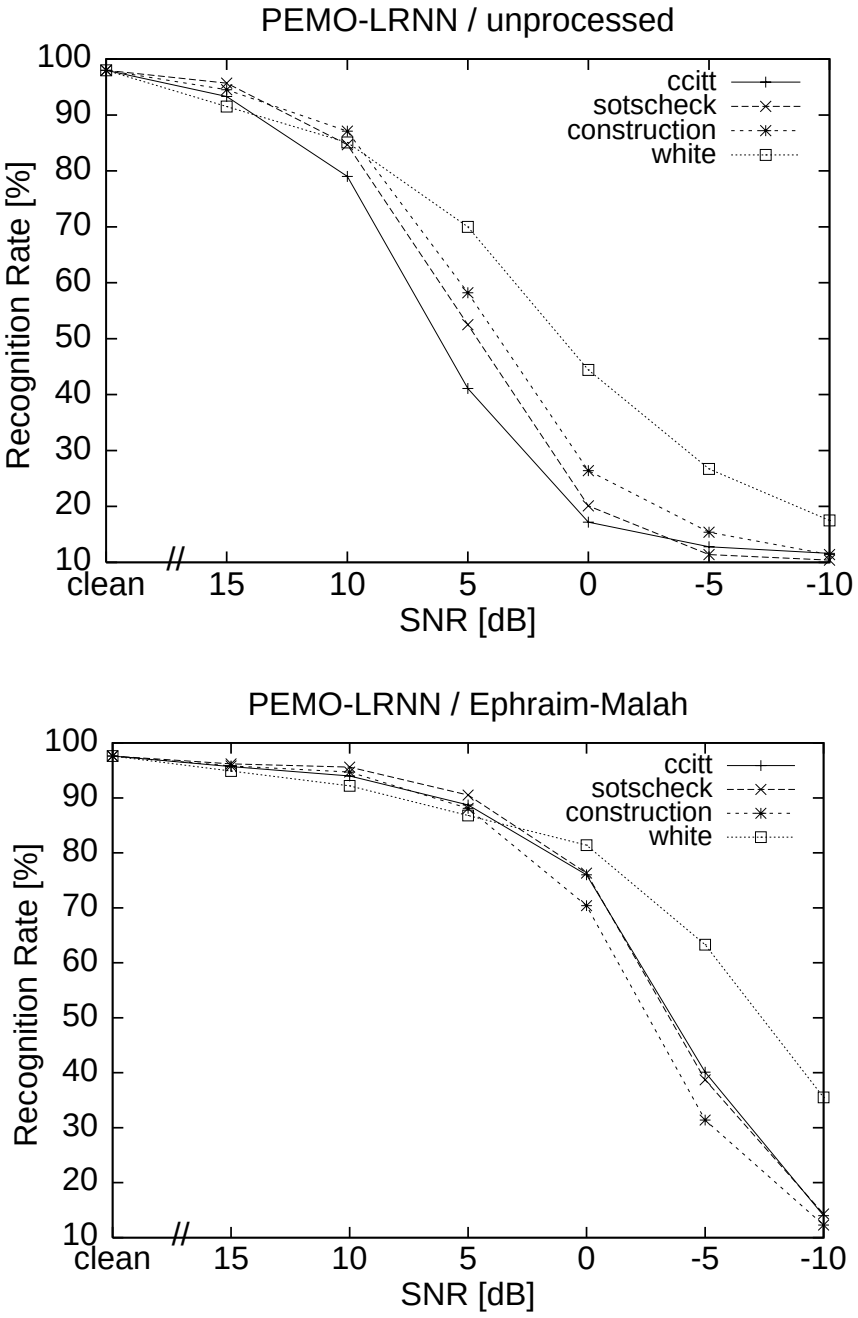


Figure 2.7: Speaker-independent, isolated digit recognition rates as function of SNR with no speech enhancement (top) and Ephraim-Malah speech enhancement (bottom) for different types of noise.

stationary nature. This effect is not unexpected, as the noise reduction scheme assumes temporarily stationary noise while speech is active.

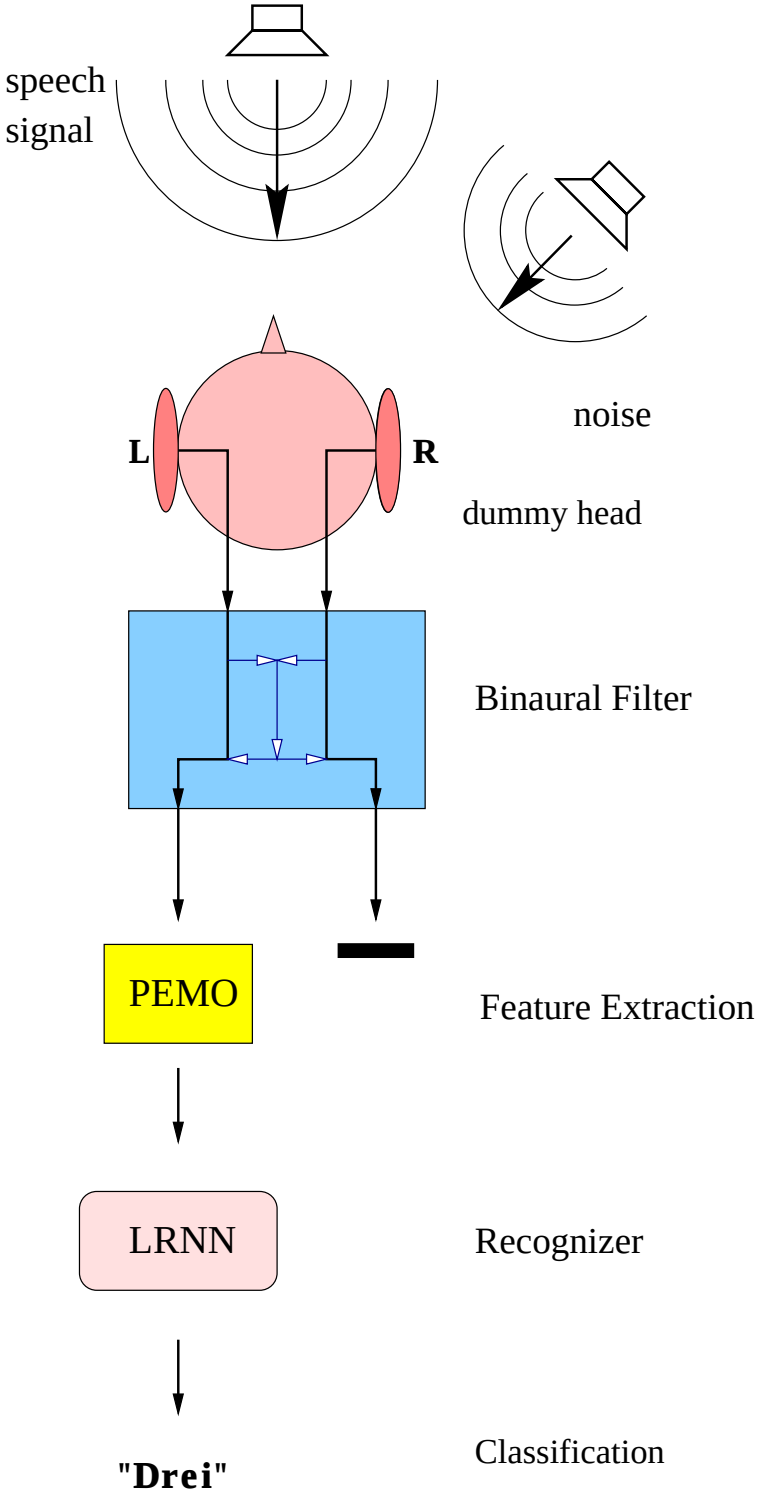


Figure 2.8: Setup of binaural isolated digit recognition experiment. Speech signals and noise were recorded separately with the same setup of Oldenburg dummy head and loudspeaker configuration. The signals from different source locations were then mixed, processed by the binaural filter and finally the left channel used for feature extraction and classification.

2.5 Digit Recognition Experiments with Binaural Speech Enhancement and PEMO Front End

While single-channel noise suppression is based on the assumption that the noise signal is stationary, multi-channel noise reduction methods, in theory, allow for a separation of different sound sources based on spatial direction alone. In addition, unwanted reverberation can be suppressed, especially when taking advantage of the directional characteristics of a human (or dummy) head. In this section, a directional filter algorithm is examined as a pre-processing step for the digit recognition system described above.

2.5.1 Setup

A two-channel algorithm for the use in binaural digital hearing aids has been proposed by Kollmeier et al. (1993), Peissig (1993) and Wittkop et al. (1997). Differences in amplitude and phase between left and right input channel frequency components are used for a directional filter. Also, the interaural coherence function serves as a basis for dereverberation. A third component has been added later (Wittkop et al., 1999) for suppression of single jammer sources. The effect on speech intelligibility has been evaluated using audiometric sentence tests (Wittkop et al., 1999). While the suppression of noise was clearly audible in informal listening tests, no significant increase of speech intelligibility could be found averaged over a number of hearing-impaired subjects. As for the monaural algorithm, however, the subjective personal preferences tended towards the filtered signals.

In a previous study virtual acoustics was used to examine the usefulness of this algorithm in the field of ASR (Kleinschmidt et al., 1998). A significant increase in recognition performance could be observed. However, to evaluate the combination of binaural filtering and the PEMO/LRNN isolated digit recognition system in more realistic conditions, actual recordings were taken as training and test data. The experimental setup is shown in Figure 2.8. The speech signals from the ZIFKOM corpus and different noise signals were re-recorded, both in an anechoic chamber and in a moderately reverberant seminar room (average reverberation time of 0.5s). The same loudspeaker was placed at different azimuth angles 2.5m away from the

Oldenburg dummy head on the horizontal plane. The signals from the built in microphones were directly recorded on hard disk. Later on, speech and noise signals were mixed at different SNRs and azimuthal directions and processed by the binaural algorithm. Finally, the left output channel underwent PEMO feature extraction and LRNN classification.

The SNR was not easy to determine because speech and noise sources were placed at different azimuthal angles relative to the dummy head. Calculating the RMS values at the sources would have meant to ignore the effect of the dummy head related transfer function and required a high technical effort when recording the signals. Instead, the RMS values were calculated using speech and noise arriving from frontal sources at 0 degrees azimuth and the corresponding gain was applied to the lateral noise recordings. In all cases the speech source was situated in front of the dummy head, while the noise source was located at different angles to the right. The LRNN training was always carried out on clean speech data, which was recorded and filtered the same way as the test corpus, i.e. the reverberant test data was evaluated using a LRNN trained on reverberant training data.

2.5.2 Results

The speaker-independent recognition results for isolated German digits recorded in the anechoic chamber are shown in Figure 2.9. When applying the binaural filtering algorithm the error in CCITT noise (30 and 60 degree) drops significantly. The effect is less pronounced for a jammer source at 30 degree azimuth and very low SNR levels. A possible explanation might be the tuning of the directional filter in this set of experiments to no attenuation between 0 and 20 degrees and maximum attenuation for all sources located at directions over 40 degrees. The maximum gain in recognition performance (60 degree) was about 60 % absolute at 0dB SNR, which corresponds to an effective gain in SNR of approximately 10dB at 90 % level. As expected, the directional filter yields no improvement in the 0 degree case, where speech and noise source are not spatially separated. As no negative effect can be observed either, possible degradations of the speech signal by artifacts of the processing are not 'noticed' by the recognition system. Furthermore, in the case of clean test data the error rate has not changed significantly by applying the binaural filtering (see Table 2.2).

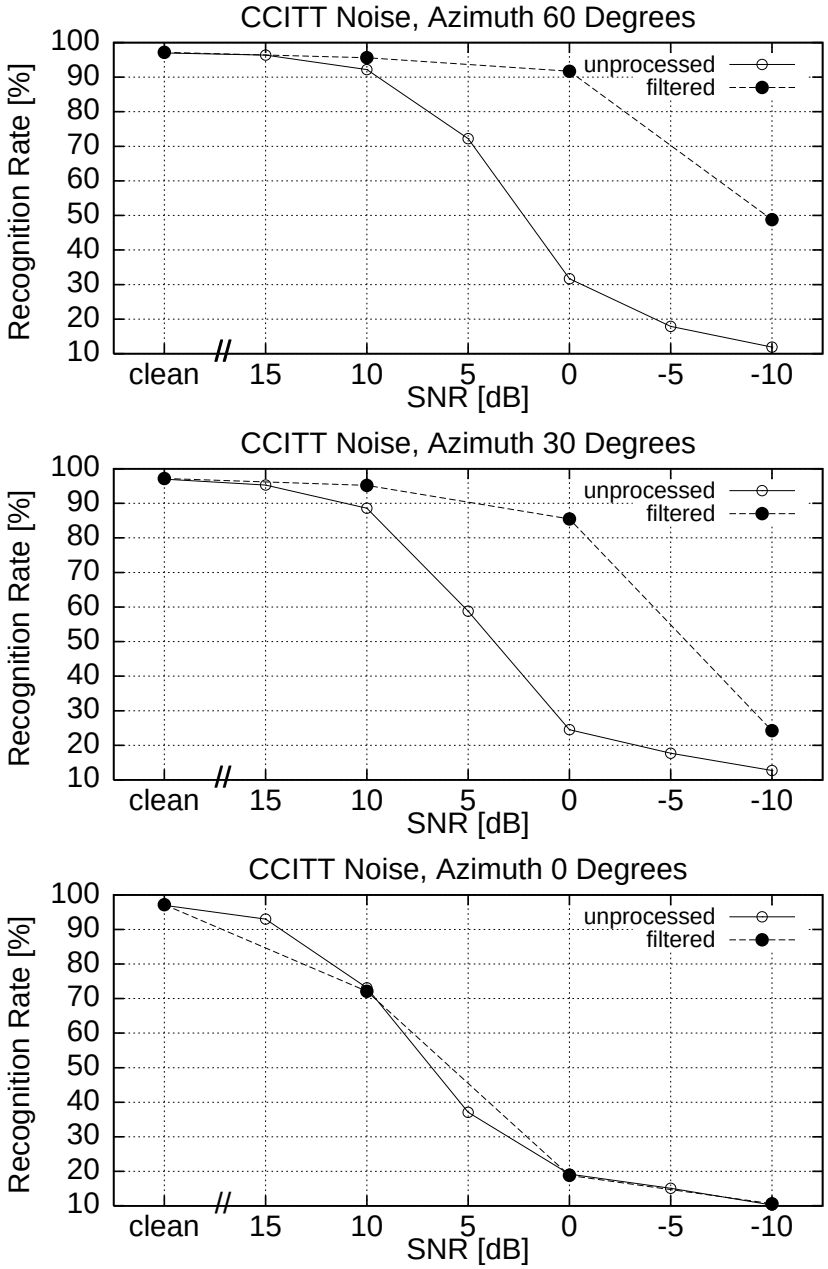


Figure 2.9: Speaker-independent, isolated digit recognition rates in CCITT noise as function of SNR in anechoic condition for different angles of noise source location.

This is also true for the experiments in reverberant environment (Figure 2.10). In this case, the overall performance is worse than in the anechoic chamber. The PEMO/LRNN recognition system is affected by reverberant conditions. Even the use of training data recorded in the same room yields higher error rates for clean test data than in the anechoic case (see Table 2.2). Moreover, the binaural filter seems to be less effective under reverberant conditions, resulting in a less pronounced enhancement of recognition

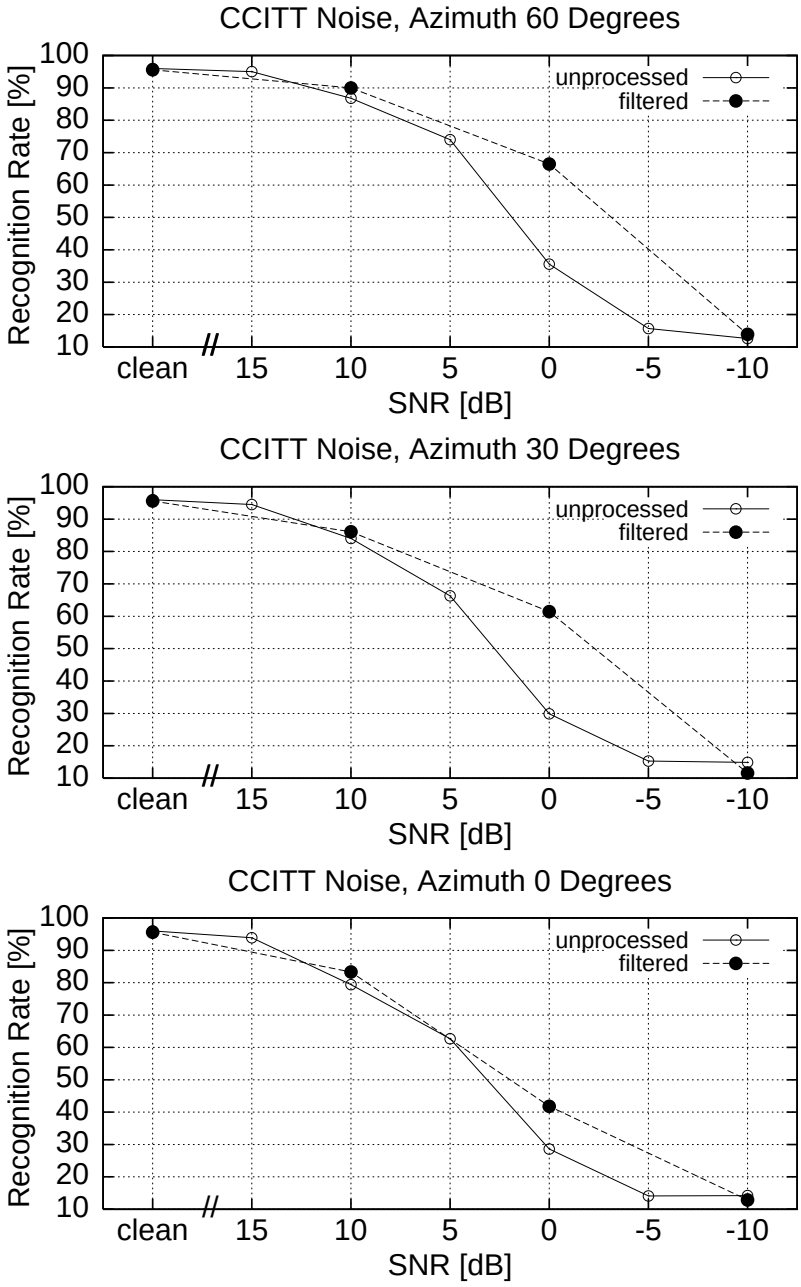


Figure 2.10: Speaker-independent, isolated digit recognition rates in CCITT noise as function of SNR in reverberant condition for different angles of noise source location.

rates for CCITT noise at 30 and 60 degree azimuth. This observed effect coincides with the results from speech intelligibility tests (Wittkop et al., 1997), where improvements on speech perception thresholds were found only in non-reverberant conditions and a small number of jammer sources.

Table 2.2: Speaker-independent isolated word recognition rate in % on clean test data for different reverberant environments with binaural filtering and dereverberation and no processing.

	without processing	with binaural filtering
anechoic chamber	97.3	97.2
seminar room	96.0	95.6

2.6 Direct Comparison of Monaural and Binaural Speech Enhancement Methods

It can be concluded from the above experiments that monaural and binaural noise reduction algorithms both have the capability to significantly increase the robustness of the PEMO/LRNN isolated digit recognition system. However, a direct comparison is still missing as the experimental setups and especially the SNR calculations were not directly comparable due to the filtering effect of the dummy head. In addition, the evaluated noise signals did not include background speech as interfering sources. Therefore, a third set of experiments was performed and will be described in this section.

2.6.1 Setup

For the following experiments the binaural setup (see Section 2.5 and Figure 2.8) has been used. In some cases the CCITT noise has been replaced by babble noise[c], which was recorded in a cafeteria. For speech enhancement, either the binaural filter or the monaural Ephraim-Malah scheme were applied. The monaural algorithm was only applied to the left channel of the recording. Again, the experiments were carried out in anechoic and in moderately reverberant surroundings using an LRNN trained on clean, unprocessed speech recorded in anechoic or reverberant conditions, respectively.

[c]NOISEX database, see Varga et al. (1992)

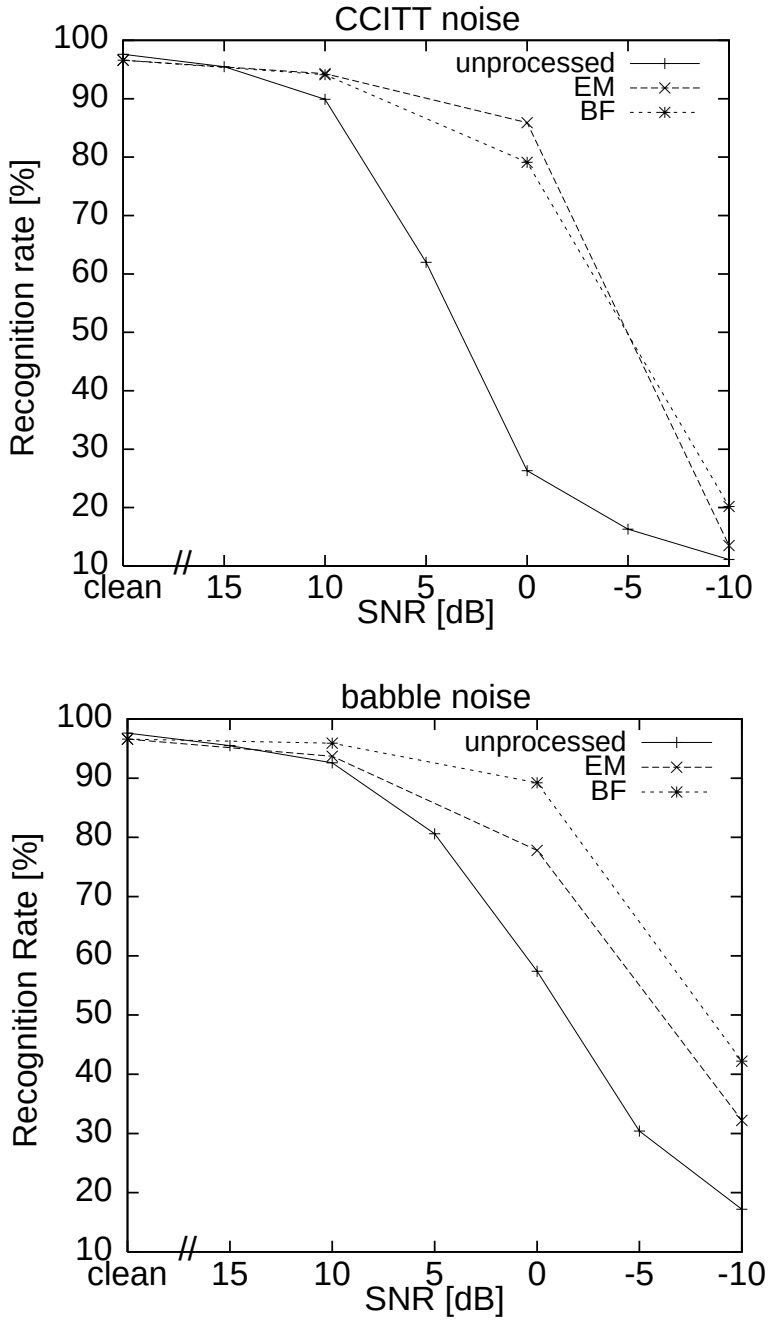


Figure 2.11: Speaker-independent, isolated digit recognition rates in CCITT noise (top) and babble noise (bottom) as functions of SNR in anechoic conditions for Ephraim-Malah (EM) and binaural filter (BF) speech enhancement and PEMO/LRNN recognition system. Speech and noise source were separated by a 30 degree azimuthal difference.

2.6.2 Results

The recognition performance of the PEMO/LRNN digit recognition system with binaural filter (BF), Ephraim-Malah (EM) and no processing for

the anechoic chamber recordings are presented in Figure 2.11. Both algorithms yield a significant improvement in robustness in CCITT noise (top) of about 60 % in absolute at 0dB SNR or an effective gain of 10dB SNR at 90 % level. The gain of performance by the two algorithms is of similar size, the Ephraim-Malah scheme being slightly superior. In contrast, the binaural filter is far more successful in the suppression of babble noise (bottom) when speech and noise source are spatially separated. In the displayed case of 30 degree azimuthal angle between speech and noise source the binaural filter leads to a 30 % gain absolute at 0dB SNR compared to 20 % with the monaural filter. As expected, the modulated characteristic of babble noise is a bigger problem for the single-channel speech enhancement. Still monaural schemes like Ephraim-Malah's have their advantages, for example in the case of 0 degree azimuthal difference between speech and noise source, for which binaural filtering is useless.

The classification results in the moderately reverberant condition are plotted in Figure 2.12. As mentioned before, the error rate for clean test data is in all cases higher than in anechoic surroundings. This indicates that the recognition system itself might be disturbed by reverberation, even though the LRNN has been trained on reverberant data. In CCITT noise (top) the monaural and the binaural algorithm yield a significant improvement in recognition performance over the unprocessed alternative. In contrast to the recognition rates obtained in the anechoic chamber, the binaural filter is less effective than the Ephraim-Malah algorithm, scoring a gain of 30 % absolute at 0dB SNR compared to 45 %. For a quasi-stationary signal like CCITT noise, reverberation is merely a change of spectral characteristics, which is no problem for the Ephraim-Malah algorithm. In contrast to that, the performance of the directional filter and noise source canceler is negatively affected by higher degree of diffusiveness.

In additive babble noise (bottom), both algorithms yield similar improvement of recognition performance under reverberant conditions, but the gain in recognition rate is much smaller than in anechoic surroundings with 10 % absolute at 0dB SNR compared to 20 and 30 %. The advantages and disadvantages of monaural and binaural noise reduction schemes, e.g. nonstationarity of the noise signal and reverberation, seem to affect the speech enhancement for ASR systems to a comparable degree. The consecutive application of both algorithms might lead to further synergetic effects and will be evaluated in future studies.

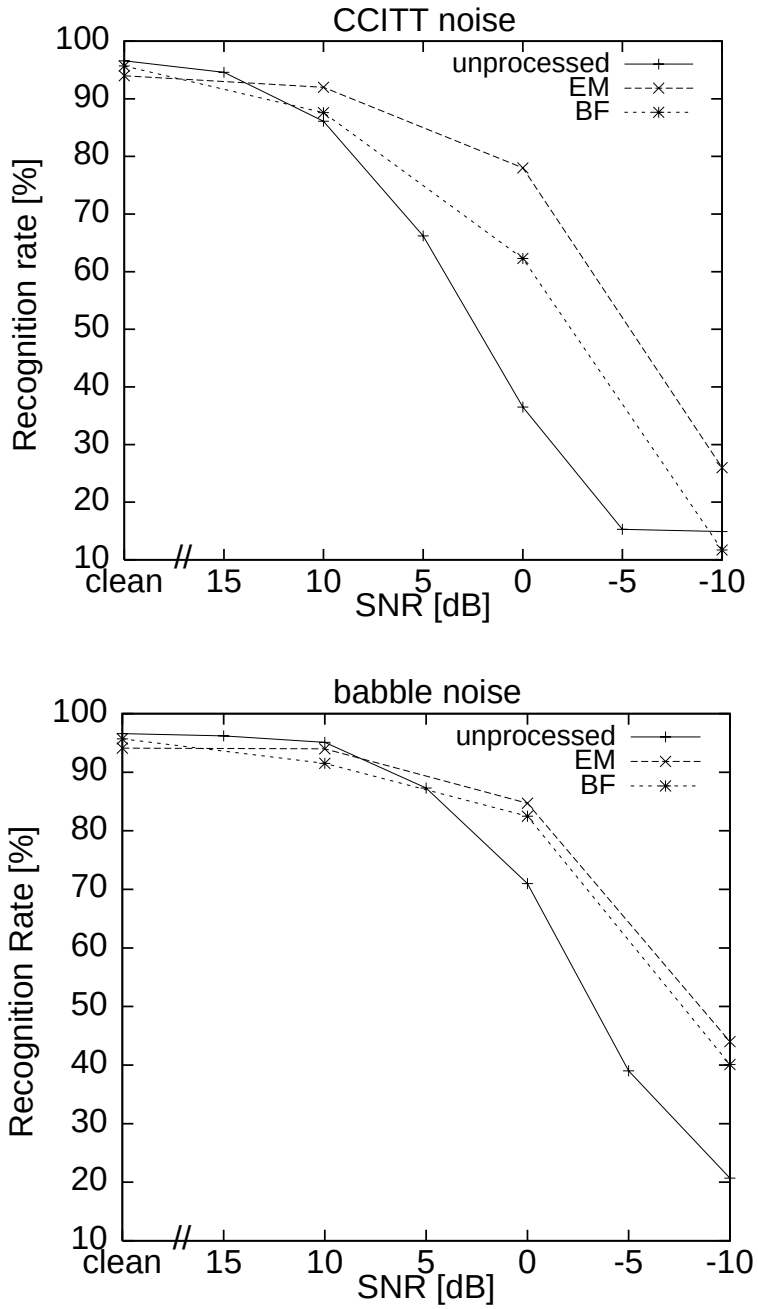


Figure 2.12: Speaker-independent, isolated digit recognition rates in CCITT noise (top) and babble noise (bottom) as function of SNR in reverberant conditions for Ephraim-Malah (EM) and binaural filter (BF) speech enhancement and PEMO/LRNN recognition system.

2.7 Discussion

The auditory model, which was originally developed for predicting human performance in psychoacoustical masking experiments, shows promising

results when applied as a front end for ASR. The intention of this quantitative model of auditory processing is to transform an incoming sound waveform into its internal representation. Rather than trying to model each physiological detail of auditory processing, the approach is to focus on the effective signal processing in the auditory system which uses as little physiological assumptions and physical parameters as necessary, while still predicting as many psychoacoustical aspects and effects as possible. A recent study (Tchorz and Kollmeier, 1999a) focuses on the amount that each processing stage of PEMO contributes to the robust representation of speech. The results show that the adaptive compression stage is of major importance in this task. The nonlinear adaptation loops yield an enhancement of changes in the input signal and suppression of steady-state portions. When the adaptation stage of PEMO is replaced by a static logarithmic compression of the amplitude in each frequency band (as in common bank-of-filters front ends), the recognition rates in quiet were high but dropped rapidly when the test material was distorted with additive noise (see Figure 2.3 on page 30). Another processing step which contributes to robust recognition is the low pass filter which smoothes the fluctuations in each frequency band after dynamic compression. The filter leads to a band pass characteristic of the amplitude modulation transfer function of the model: slow modulations are suppressed by the adaptation loops, fast modulations by the filter. The maximum in the modulation transfer function in the original model is at approximately 6Hz. Shifting the maximum to 4Hz by modified low pass filtering further enhances robustness of ASR in noise. This might be explained by the better correspondence with the average modulation spectrum of speech, which has its maximum at around 4Hz. Fast fluctuations in the input signal which are not likely to origin from speech are better suppressed with a modified low pass filter. A more detailed study by Kanedera et al. (1999) on modulation processing of ASR front ends supports this hypothesis.

A further improvement of the robustness of ASR systems can be achieved by applying monaural and binaural noise suppression schemes to the disturbed input signal. The PEMO front end has shown to yield robust recognition performance not only against additive noise but also against possible distortions and artifacts introduced by the noise reduction algorithms. It seems to work especially well when combined with speech enhancement methods originating from digital hearing aid technology. Although (hearing-impaired) human listeners have not been shown to gain a significant advantage from speech enhancement schemes in terms of speech

intelligibility, the PEMO/LRNN recognition system obviously benefits to a large degree. This may be caused by two factors: a) the range of SNRs necessary to obtain 50 % intelligibility in normal and most hearing impaired listeners is still lower than the SNR required to achieve a 50 % recognition rate for the ASR systems tested here. Since the performance of noise reduction schemes usually degrade with decreasing SNR, the higher gain in 'intelligibility' for ASR applications might be due to this difference in original SNR level employed. b) the highly efficient cognitive system of normal and hearing-impaired human listeners is able to compensate for unfavorable SNR conditions by decomposing the incoming sound image into desired speech and undesired background noise. The ease of listening tests and subjective preferences of human subjects indicate that a major cognitive effort is needed for the human brain to make the noise suppression algorithms redundant. This ability is not included in the usual construction of ASR systems. Hence, they have to rely on an appropriate (acoustical) pre-processing to obtain a high enough SNR.

For the purpose of robust ASR the results are very promising. Besides the major increase of recognition performance which is equivalent to an effective gain in SNR by 5 to 10dB in most cases, it is very important that the error rate for clean test data did not change significantly. The positive effect of the noise reduction algorithms was found for all examined types of noise as well as for all SNR and spatial configurations. The exceptions to this general trend were expected, e.g. no increase with binaural filtering at 0 degree azimuth between speech and noise source, or a limited effect of the Ephraim-Malah algorithm on highly modulated speech noise (babble).

It should be noticed here that in this paper the ASR systems were always trained on *clean* training data. It can not be concluded that the advantage of the PEMO front end and especially speech enhancement methods still holds when training is performed differently, e.g. using clean *and noisy* input data. However, since the type and level of disturbing noise in practical conditions is generally not known a priori, it is expected that the advantage of the PEMO front end demonstrated here may still hold in practical applications when untrained noise is encountered.

As a major problem the degraded performance in reverberant environments remains. The PEMO/LRNN recognition system shows no optimal performance even when trained with reverberation, i.e., training on data recorded in the same room as the test data. Also the binaural filter algorithm is

affected by a high degree of diffusiveness in the input signals, while the monaural algorithm seems to work similarly well in all surroundings.

The intention of this paper is to demonstrate the usefulness of speech enhancement techniques to already robust PEMO based ASR systems which had been tested successfully against other front ends in the past. Speech enhancement techniques which were originally designed to increase speech intelligibility of (hearing impaired) human listeners were combined with auditory based feature extraction. There is a large variety of other techniques which can partly be applied after feature extraction or are based on a modified classifier. Combining these different approaches to robust speech recognition might yield recognition systems even more robust towards additive and convolutive noise. The idea behind this paper is to show that speech enhancement combined with auditory based feature extraction might be a promising candidate to play an important role in that task.

2.8 Outlook

To further evaluate the usefulness of PEMO as front end for ASR systems, experiments with extended vocabulary (more than only 10 digits) or based on sub-word units are necessary. When transformed to PEMO-CEP features, the PEMO internal representation of signals could be examined as a front end for standard HMM phoneme based recognition systems. In addition, the PEMO/LRNN system needs to be optimized for real-world applications in reverberant environments. This is especially important when it comes to hands-free input devices. Binaural filter algorithms in principal have the capability to reduce reverberation, yet the one applied in this study suffers more by the presence of acoustical echoes than its monaural counterpart. A combination of the binaural filter with a successive Ephraim-Malah noise reduction might result in a further synergistic improvement of performance (see Meyer and Simmer, 1997, for a combination of binaural and monaural processing.). Such a combination has been evaluated already for hearing-impaired subjects wearing hearing aids (Marzinzik et al., 1999).

To aquire the results of speech intelligibility and ease of listening tests, extensive and time-consuming experiments had to be carried out with (hearing impaired) human subjects. Although the requirements and SNR condi-

tions are somewhat different between human listener speech intelligibility tests and ASR experiments, the methods proposed here suggest an 'objective' way to evaluate noise reduction algorithms also for other types of applications, e.g., hearing aid and telecommunication technology.

Thanks to Mark Marzinzik and Thomas Wittkop for supplying their implementation of the speech enhancement algorithms and much valuable advice. Thanks also to Volker Hohmann for fruitful discussions and his support.

Klaus Kasper and Herbert Reininger, from the Institut für angewandte Physik, Universität Frankfurt, for supplying their LRNN implementation.

Part of this work was supported by Deutsche Forschungsgemeinschaft (KO 942/12-1).

CHAPTER 3

Sigma-pi Cells as Secondary Features for Isolated Digit Recognition [a]

Contents

3.1 Introduction and Summary 51
3.2 Recognition System 52
3.3 Experiments 55
3.4 Summary and Outlook 57

3.1 Introduction and Summary

One of the major problems in automatic speech recognition (ASR) is the lack of robustness in adverse acoustical conditions. In this chapter, the combination of auditor model based feature extraction and the Feature-finding Neural Network (FFNN) is evaluated. The model of auditory perception (PEMO) after Dau et al. (1996a) is an effective model designed to simulate psychoacoustical experiments. In Chapter 2 it is shown that PEMO is an especially robust front end for isolated digit recognition in combination with speech enhancement techniques. In the FFNN approach (Gramß and Strube, 1990; Gramß, 1992) new spectro-temporal features are derived from the initial representation and classification is performed by a linear neural network. The parameter sets of these secondary features are optimized based on the training corpus. The results obtained

[a] A shorter German version of this chapter appeared on pp. 382–383 in *Fortschritte der Akustik - Proceedings of DAGA 2000* by Michael Kleinschmidt and Volker Hohmann as 'Perzeptive Vorverarbeitung und automatische Selektion sekundärer Merkmale zur robusten Spracherkennung'.

by PEMO/FFNN on isolated German digits in different noise conditions indicate a more robust performance than the reference system.

Einleitung und Zusammenfassung

Für die automatische Spracherkennung ist die mangelnde Robustheit gegenüber additiven Störgeräuschen eines der ungelösten Probleme. In diesem Kapitel wird die neuartige Kombination einer perzeptiven Vorverarbeitung mit einem speziellen neuronalen Netz vorgestellt und ihre Robustheit gegenüber Störschallen evaluiert. Das Perzeptionsmodell (PEMO) nach Dau et al. (1996a) ist ein effektives Modell der auditorischen Signalverarbeitung und wurde ursprünglich zur Simulation psychoakustischer Experimente konzipiert. Es ist insbesondere in Kombination mit Methoden der Störgeräuschunterdrückung eine robuste Vorverarbeitung für die Einzelworterkennung im Störgeräusch (Chapter 2). In diesem Kapitel wird die Kombination des PEMO mit dem *Feature-finding Neural Network* nach (Gramß and Strube, 1990; Gramß, 1992) evaluiert. Dabei werden aus den primären (PEMO) Merkmalen zunächst neue temporale und spektrale Merkmale extrahiert und diese anschließend mit Hilfe eines linearen neuronalen Netzes klassifiziert. Diese sekundären Merkmale werden während der Trainingsphase automatisch optimiert. Die vorgestellten Ergebnisse zur sprecherunabhängigen Klassifikation isolierter deutscher Ziffern in unterschiedlichen Störschallsituationen weisen auf eine Verbesserung der Robustheit im Vergleich zu anderen Erkennungssystemen hin.

3.2 Recognition System

For robust feature extraction the model of auditory perception after Dau et al. (1996a) is used. This is combined with the Feature-finding Neural Network (FFNN) after Gramß (1992). Both parts are derived from neurophysiological and psychoacoustical knowledge, but have neven been investigated in this combination before.

3.2.1 Perceptual Feature Extraction

The model of auditory perception (PEMO) after Dau et al. (1996a) is an effective model of the signal processing in the peripheral auditory system. It quantitatively simulates the response of human subjects in a number of psychoacoustical experiments such as spectral and temporal masking. PEMO converts the incoming waveform into an *internal representation*, which has been shown to be a robust feature for ASR (Tchorz and Kollmeier, 1999a). This is especially true, when PEMO is used in combination with a neural network classifier for isolated word recognition in additive noise (Tchorz et al., 1997; Kasper et al., 1997; Tchorz et al., 1999). The PEMO model consists of an initial pre-emphasis and a subsequent gammatone filterbank for frequency decomposition (see Figure 2.1 on page 25). This is followed by channelwise half-way rectification and envelope extraction. Five nonlinear adaptation loops perform compression. Together with a subsequent modulation low pass they model adaptation effects. The thereby obtained internal representation is averaged over 10ms and then forms the feature vector sequence. In this application, 15 filters are used with width and spacing of two ERB. The center frequencies are between 50 and 7000 Hz. The filter width of two ERB instead of one is more in accordance with *articulation bands*, which are important for speech perception (Fletcher, 1953; Allen, 1994), than with *critical bands*.

3.2.2 Secondary Feature Extraction

Gramß and Strube (1990) proposed the Feature-finding Neural Network (FFNN) as a classifier for ASR. They consist of a stage for the extraction of special *secondary features* from the original, *primary feature* vector sequence. The secondary features are then used as input to a linear neural network classifier. The secondary features used here, are called sigma-pi cells and derived from the primary feature vectors $\vec{m}(t)$ as follows:

$$x(f,t,f_0,t_0,\Delta f,\Delta t) = m(f,t) \cdot \sum_{f'=0}^{\Delta f-1} \sum_{t'=0}^{\Delta t-1} m(f+f_0+f',t+t_0+t')$$

A small and a large window are used to extract specific parts of the primary feature matrix or sequence. The total sum over the large window is then multiplied by the small window value (hence the name sigma-pi). t and

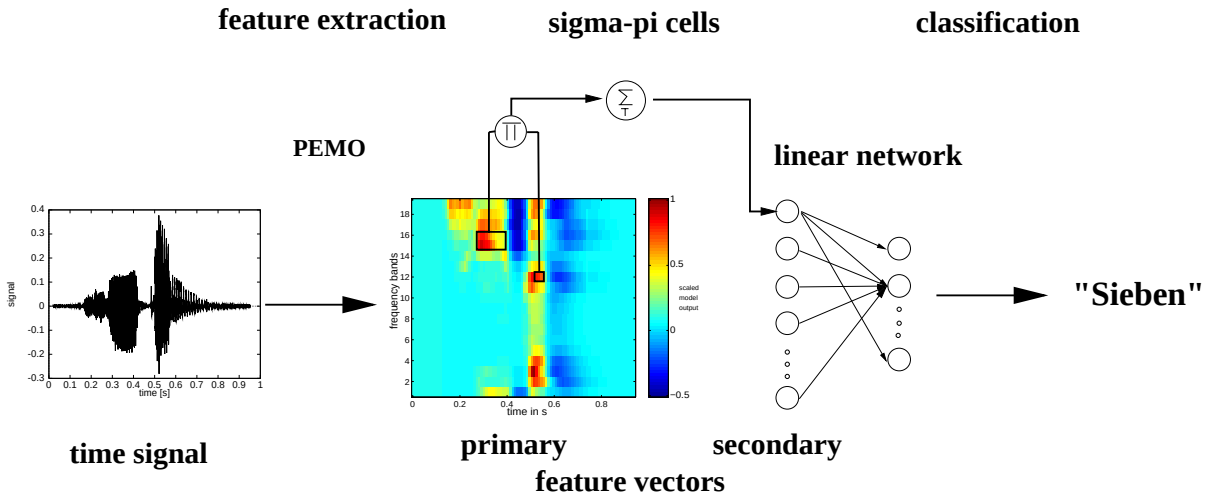


Figure 3.1: Scheme of the PEMO/FFNN recognition system. Auditory model processing (PEMO) yields a spectro-temporal representation of primary feature vectors. Sigma-pi cells extract secondary features from that representation and serve as input to the linear classifier.

f denote time and frequency axis, respectively, while f_0 and t_0 denote the distance of the two windows and Δf, Δt the size of the larger window. For isolated word recognition, the resulting secondary feature values $x(f,t)$ are summed over the whole time of the utterance. This results in only one secondary feature vector $\vec{x}$ for the complete utterance.

Gramß and Strube (1990) had already motivated the sigma-pi approach by findings in physiology and psychoacoustics. The resemblence of certain sigma-pi cells with recent studies on early auditory features in psychoacoustics is especially interesting. Kaernbach (2000) found very similar features by reverse correlation with masking experiments for periodic noise stimuli.

3.2.3 Feature Set Optimization

Due to the simple linear classifier, the optimal weight matrix $\mathbf{W}$ may be found analytically. Let N be the number of secondary features, M the number of classes and P the number of examples in the training set ($M \times$ # speakers $\times$ examples per class and speaker). $\mathbf{X}$ then contains a secondary feature vector of an example in each column and the classification problem is stated as:

$$\begin{array}{ccccc} \tilde{\mathbf{Y}} & = & \mathbf{W} & \cdot & \mathbf{X} \\ (M \times P) & = & (M \times N) & \cdot & (N \times P) \end{array}$$

After Gramß (1992) the optimal weight matrix $\mathbf{W}$ is calculated by

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^{+} \text{ , wobei } \mathbf{X}^{+} = \mathbf{X}^{T}\mathbf{\Psi} \text{ , } \mathbf{\Psi} = (\mathbf{X}\mathbf{X}^{T})^{-1}$$

if the rank of $\mathbf{X}$ is reaches the maximum value and $P \geq N$. $\mathbf{X}^{+}$ is called the *pseudoinverse* of $\mathbf{X}$. The feature selection problem may be solved automatically as the computational effort is relatively small for a certain type and number of secondary features. The *substitution rule* after Gramß (1992) is used here. It starts with N randomly chosen features and each iteration the least relevant feature is discarded and replaced by a randomly drawn new one. The relevance R_i of feature/sigma-pi cell i is derived from the Euclidean error $E = \|\mathbf{Y} - \tilde{\mathbf{W}}\mathbf{X}\|^2$ as follows:

$$R_i = \Delta E_i = E(\text{without feature i}) - E(\text{with feature i})$$

In the experiments below, the optimization was stopped after 500 iterations.

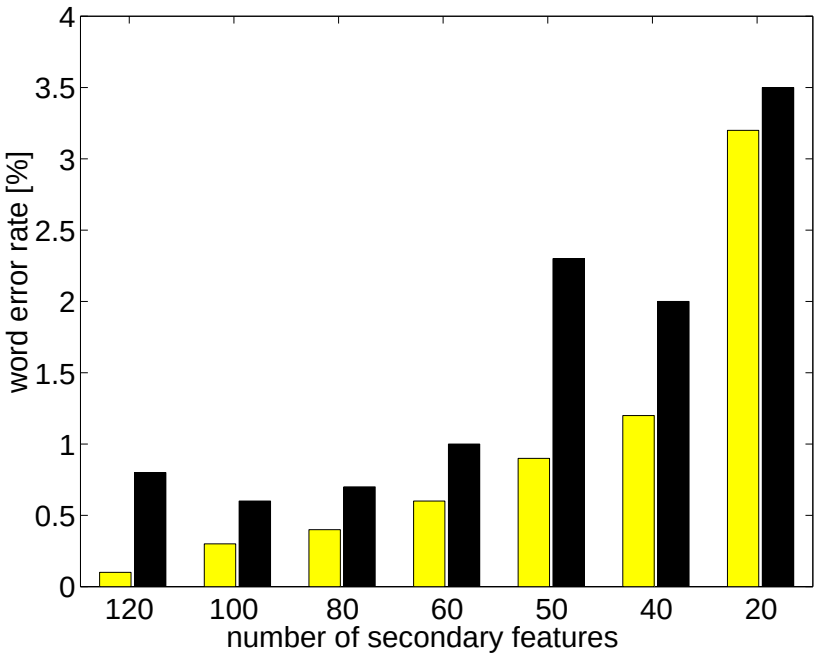
3.3 Experiments

The following experiments are restricted to isolated word recognition. Parts of the ZIFKOM corpus (Deutsche Telekom) were used. Ten German digits were recorded once for each of the 100 female and 100 male speakers. Training and test set are disjoint with respect to the speakers and consist of 1000 utterances each. The classification is speaker independent.

3.3.1 Optimal Number and Types of Features

It was evaluated how many secondary features are necessary to yield an acceptable recognition score. Figure 3.2 shows the error rates for clean training and test material, as well as for speech simulating CCITT noise (CCITT G.227) at 10dB SNR depending on the number of secondary features. The word error rate (WER) for clean test data increases significantly with less than 60 features. In the noisy condition, the WER shows an increase already at below 80 features. $N = 80$ features seems to be a good compromise between performance and computational effort as the size of matrix $\mathbf{X}\mathbf{X}^{\mathbf{T}}$ increases with N^2. Therefore, in the following experiments always 80 features were used.

a) clean training and test (dark bars) data:



b) CCITT noise added at 10dB SNR:

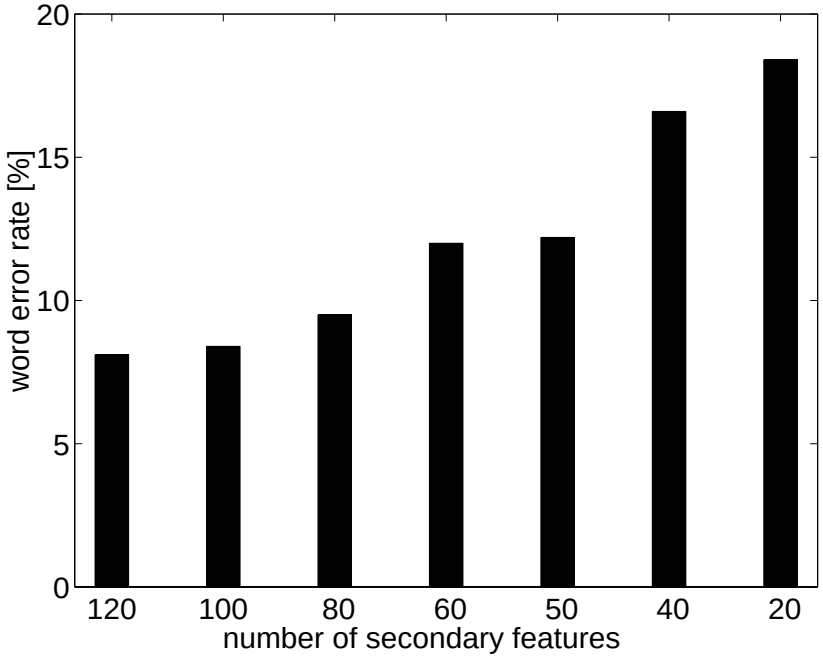


Figure 3.2: Word error rates in % depending on the number of secondary features.

During training, the free parameters of a new secondary feature were drawn randomly from a equal probability distribution across the following allowed ranges: : $f \in [1, 15]$, $f_0 \in [-14, 14]$ (all frequency channels) and $t_0 \in [0, 30]$ (equals 300 ms) as well as $\Delta f \in [1, 5]$ and $\Delta t \in [1, 5]$. This results in a num-

ber of approximately $2 \cdot 10^6$ allowed parameter combinations. Another experiment showed that the classification performance only slightly degrades if the extension of the larger window is restricted to ($\Delta f = \Delta t = 1$). If the distance of the two windows in time or frequency is further constraint to be zero, the error rate quadruples in the former and doubles in the latter case (CCITT at 10dB SNR). This indicates the necessity of spectro-temporal integration for some of the features in this application.

3.3.2 Robustness

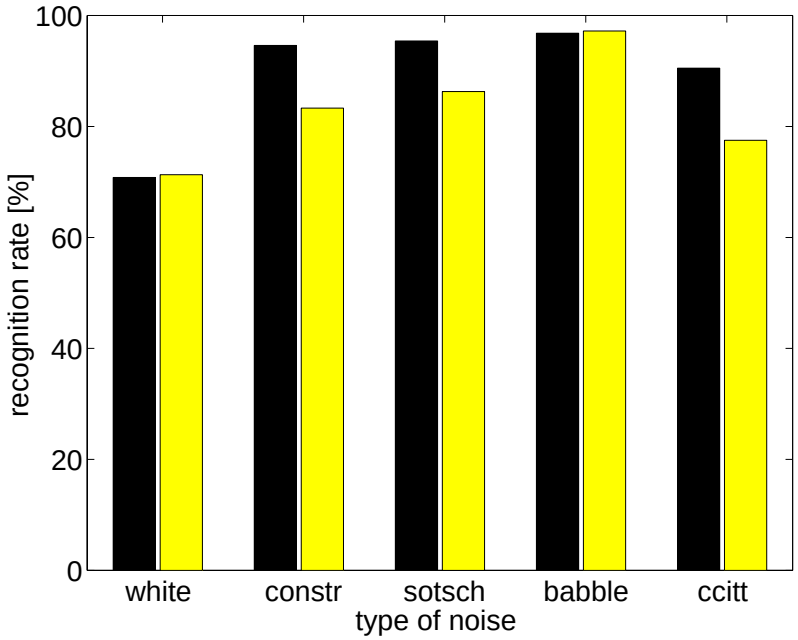
Another experiment further examines the robustness of the PEMO/FFNN system and compares the performance to reference system comprised of PEMO and a Locally-recurrent Neural Network (LRNN) as used by Kasper et al. (1995, 1997). The PEMO/LRNN system has been proven to be relatively robust in additive noise for isolated word recognition tasks (Tchorz et al., 1997; Tchorz and Kollmeier, 1999a, Chapter 2). Besides CCITT noise, also white Gaussian noise (WHITE), another speech simulating noise (SOTSCH, as in Kollmeier et al., 1988) and construction site noise (CONSTR, from Siemens, 1992) and babble noise (BABBLE, from Varga et al., 1992) were used.

Figure 3.3 a) shows the word recognition scores for PEMO/LRNN and PEMO/FFNN systems in the presence of different types of additive noise at 10dB SNR. In all cases the new PEMO/FFNN system shows performance comparable or superior to the already robust reference. Figure 3.3 b) shows the word recognition scores for PEMO/LRNN and PEMO/FFNN systems in the presence of additive CCITT noise at different SNR levels. Again the PEMO/FFNN systems outperforms the reference significantly. This is true for all SNR levels.

3.4 Summary and Outlook

The introduced combination of auditory feature extraction (Dau et al., 1996a) and Feature-finding Neural Network (Gramß, 1992) exhibits a robust classification of isolated German digits in additive noise. The PEMO/FFNN system further increase the performance of the already robust PEMO/LRNN system. The WER increases when the sigma-pi cell parameter combinations are constraint to purely temporal or purely spectral

a) several types of additive noise at 10dB SNR:



b) CCITT noise at different SNR levels:

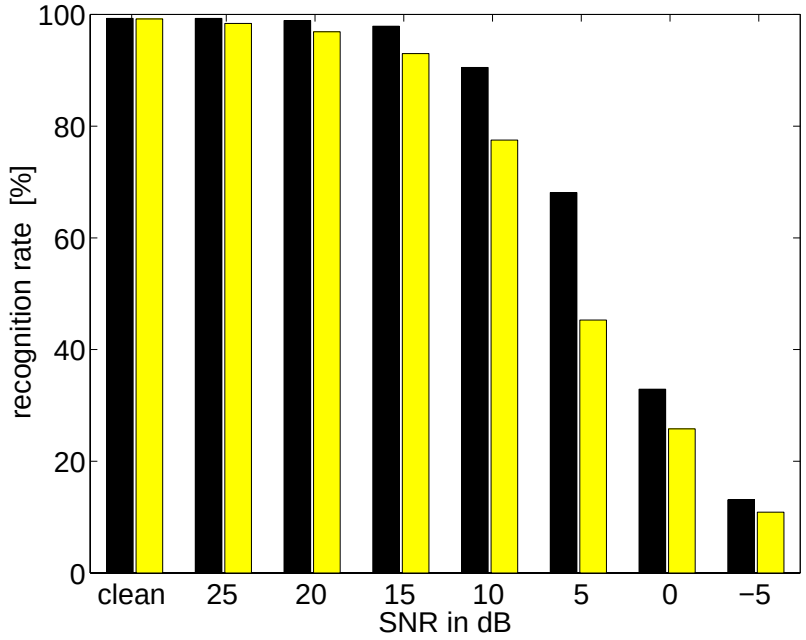


Figure 3.3: Word recognition scores for FFNN (dark bars) and LRNN (light bars).

integration. This indicates the importance of combined spectro-temporal processing for secondary feature extraction. This has to be analyzed further. Also the application of the secondary feature approach to continuous

speech recognition is to be studied, as the temporal summation of sigma-pi values cannot be used in this case.

Many thanks to Klaus Kasper and Herbert Reininger from Universität Frankfurt for providing the LRNN implementation.

chapter 4

Sigma-pi Cells as Secondary Features for Phoneme Recognition

Contents

4.1 Introduction 62
4.2 Description of Sigma-pi Cells 62
4.3 Why Using Sigma-pi Cells ? 63
4.4 Experimental Setup 64
4.5 Evaluation 68
4.6 Discussion 72

Abstract

In this chapter, it is evaluated whether sigma-pi cells may be used for phoneme classification tasks. Sigma-pi cells are second order features derived from spectro-temporal representations. Based on the output of a perception model, a large set of possible sigma-pi cells is ranked depending on their discriminative ability measured by the Fisher ratio. Also, the linear classification performance of sub-sets of all possible sigma-pi cells is presented. The results show that sigma-pi cells are principally suitable for separating different phoneme classes. A closer analysis of the parameters of those sigma-pi cells, which perform best for sub-sets of the corpus, reveals some insight about phoneme variability.

Zusammenfassung

In diesem Kapitel wird untersucht, inwieweit sich Sigma-Pi Zellen zur Phonemklassifikation eignen. Sigma-Pi Zellen sind Merkmale zweiter Ordnung, die von spektro-temporalen Mustern abgeleitet werden. In diesem Fall wird ein Perzeptionsmodell verwendet. Eine Liste von Sigma-Pi Merkmalen wird anhand ihrer Trennfähigkeit (Fisher ratio) bewertet. Zudem wird ein Set von Merkmalen zur linearen Klassifikation verwendet. Die Ergebnisse zeigen, das Sigma-Pi Zellen prinzipiell zur Klassifikation von Phonemen geeignet sind. Eine genauere Analyse der Parameter dieser optimierten Zusammenstellung von Sigma-Pi Zellen gibt Einblicke in die Variabilität der Phonemrepräsentation.

4.1 Introduction

In 1990, sigma-pi cells have been proposed to be used as secondary features based on critical band spectrograms for isolated word recognition (Gramß and Strube, 1990). Motivated by the striking resemblance of sigma-pi cell characteristics and early auditory features derived from psychoacoustical and physiological data, experiments are now carried out to use sigma-pi cells for secondary feature extraction on primary feature vectors derived from a perception model. While this approach has been successfully applied to word recognition tasks and is documented to increase the robustness in additive noise (Chapter 3), it is now to be ascertained whether these psychoacoustically motivated features are suitable for phoneme classification. In this paper two question are investigated:

1. Does this combination of perception model and sigma-pi cells yield features of sufficient discriminative ability ?

2. Which types of sigma-pi cells perform best for phoneme recognition tasks ?

4.2 Description of Sigma-pi Cells

Sigma-pi cells are known as second order elements from artificial neural network theory. The term describes certain neurons in which the weighted

outputs from two or more earlier layer neurons are multiplied before summation over all input pairs. In this paper sigma-pi cells are always defined on spectro-temporal representations (primary feature vectors). The cells consist of two windows of constant distance in time and constant frequency position. The secondary features $x(t_1, f_1, t_2, f_2, S)$ are calculated from the primary feature vectors $p(t, f)$ as follows:

$$x(t_1, f_1, t_2, f_2, S) = \sum_{t=T-S}^{T+S} p(t_1 + t, f_1) \cdot p(t_2 + t, f_2) \tag{4.1}$$

with t_n and f_n as coordinates of the n-th window in time and frequency domain and $2S+1$ as the number of feature vectors to be integrated. Note that S is always integer and that the integration is always performed symmetrically forward and backward in time. T denotes the time position of the center of a given phoneme. In the Feature-finding Neural Network (FFNN) proposed by Gramß and Strube (1990); Gramß (1991, 1992) the sigma-pi cells form the first processing part of the word classifier. The time invariance problem is solved by giving one of the windows a size larger than 1 element of the spectro-temporal representation and integrating the values for each sigma-pi cell over the primary feature vectors of the complete utterance (i.e. word) to be classified. In this paper, both windows are of size 1×1 for simplicity and because larger second windows have not lead to significant change in performance in earlier studies on isolated word recognition (Chapter 3). In addition, temporal integration is only performed over $2S + 1$ feature vectors.

4.3 Why Using Sigma-pi Cells ?

Sigma-pi cells were originally proposed for ASR in order to better capture certain features of speech like formants, formant transitions, fricative onsets and (for larger units) phoneme sequences. A logical "AND" operation is performed by multiplicative combination of the two spectro-temporal windows. This corresponds to the biological counterpart of (cortical) neurons tuned to certain spectro-temporal modulation. It is well known from psychoacoustical threshold experiments that the sensitivity of human listeners to temporal and spectral modulation peaks at around 2-8 Hz and 0.25-2 cyc/oct, respectively (Chi et al., 1999). In psychoacoustical reverse correlation experiments, using short segments of semiperiodic white gaus-

sian noise as stimuli, 'early auditory features' of certain spectro-temporal shape were revealed (Kaernbach, 2000). These findings correspond well to physiological measurements of spectro-temporal receptive fields of neurons in the primary auditory cortex (deCharms et al., 1998) which often encompass different unconnected but highly localized parts of the spectrogram.

One may argue whether the resemblance of these basic elements of auditory perception to the spectro-temporal properties of speech is coincidental or not. It is clear that many approaches to robust feature extraction for ASR implicitly take advantage of it. While the approach of calculating temporal delta features in single channels relies on comb filter effects in the modulation frequency domain, RASTA processing for example applies an effective modulation bandpass which is comparable to human auditory processing (Hermansky and Morgan, 1994). More recent data derived features are on the verge of changing from pure temporal processing to spectro-temporal integration of information (Chang et al., 2000; Kajarekar et al., 2001; Somervuo, 2002). Consecutive peaks or valleys with a specific distance in any possible direction in the spectro-temporal representation may be detected by sigma-pi cells. By using second order features such as sigma-pi cells the secondary feature vectors become even sparser than the already sparse (at least in the case of the perception model) primary features allowing for easier integration over time and better linear classification.

4.4 Experimental Setup

4.4.1 Corpus

The experiments are carried out on parts of the PhonDat 2 (PD2) corpus (2nd edition) obtained from BAS (Bayerisches Archiv für Sprachsignale, Universität München) consisting of 2,483 sentences and stories read by 33 female and 29 male speakers in the training set and 425 sentences and stories read by 11 female and 8 male speakers in the test set. The phonetic transcription has been done partly automatically and partly by hand (stories). From this, 41 classes of phonemes are extracted using the phoneme boundaries provided. In average the training and test portion consists of 2,175 and 417 examples per class.

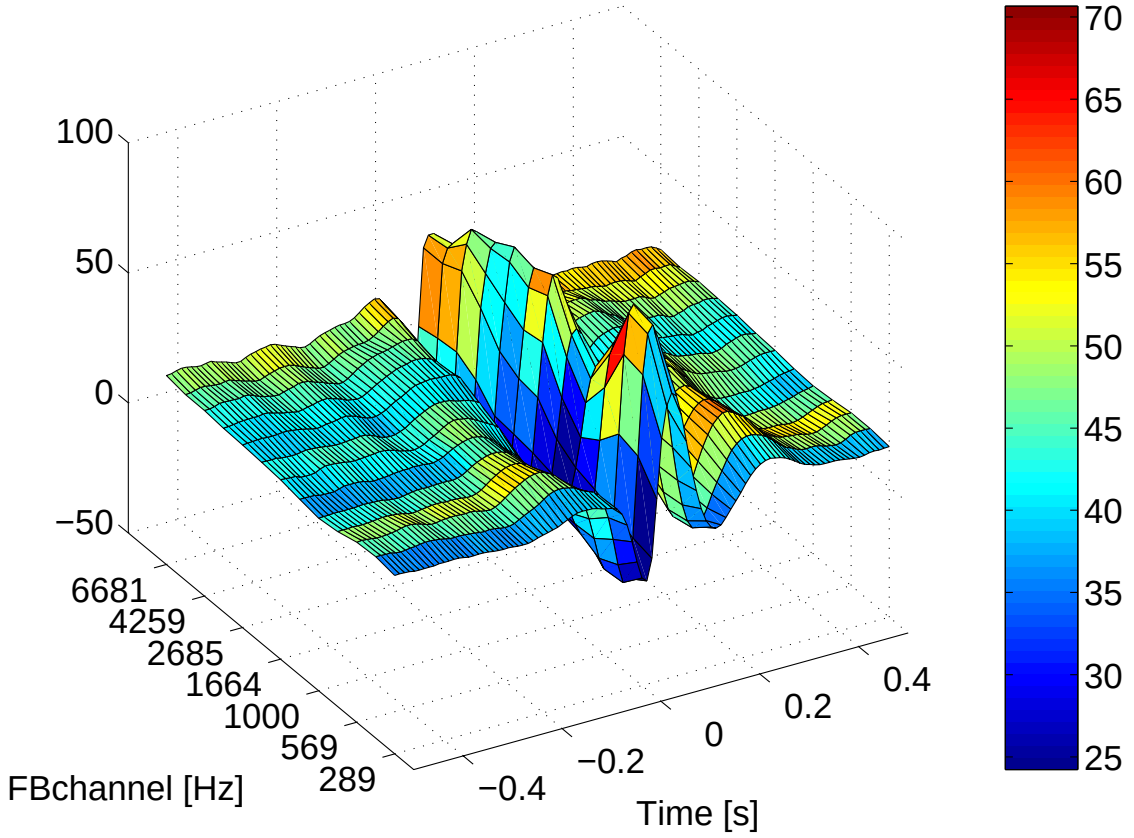


Figure 4.1: PEMO-derived spectro-temporal representation of phoneme /d/. Mean (z-axis) and standard deviation (coloring) in model units, calculated from 4394 instances of this voiced stop consonant.

4.4.2 Feature Extraction

The perception model (PEMO), used in this study, has been originally developed by Dau et al. (1996a) for quantitatively simulating psychoacoustical experiments, such as temporal and spectral masking. It has been successfully applied to robust isolated word recognition in the past (Tchorz and Kollmeier, 1999a, Chapter 2). Its major components are the peripheral gammatone filterbank and the non-linear adaption loops, which perform a log-like compression for stationary signals and emphasize onsets and offsets of the envelope (cf. Figure 2.1 on page 25). This causes a sparse coding of the input in the spectro-temporal domain. In combination with the final first order lowpass the model exhibits a modulation bandpass characteristic with a best frequency of about 4Hz. In this study, a slightly modified version of PEMO is used, which consists of an additional pre-emphasis (1st order Butterworth highpass, cutoff at 7kHz) and a modified peripheral filterbank. Filter width and spacing are increased from one to two ERB to

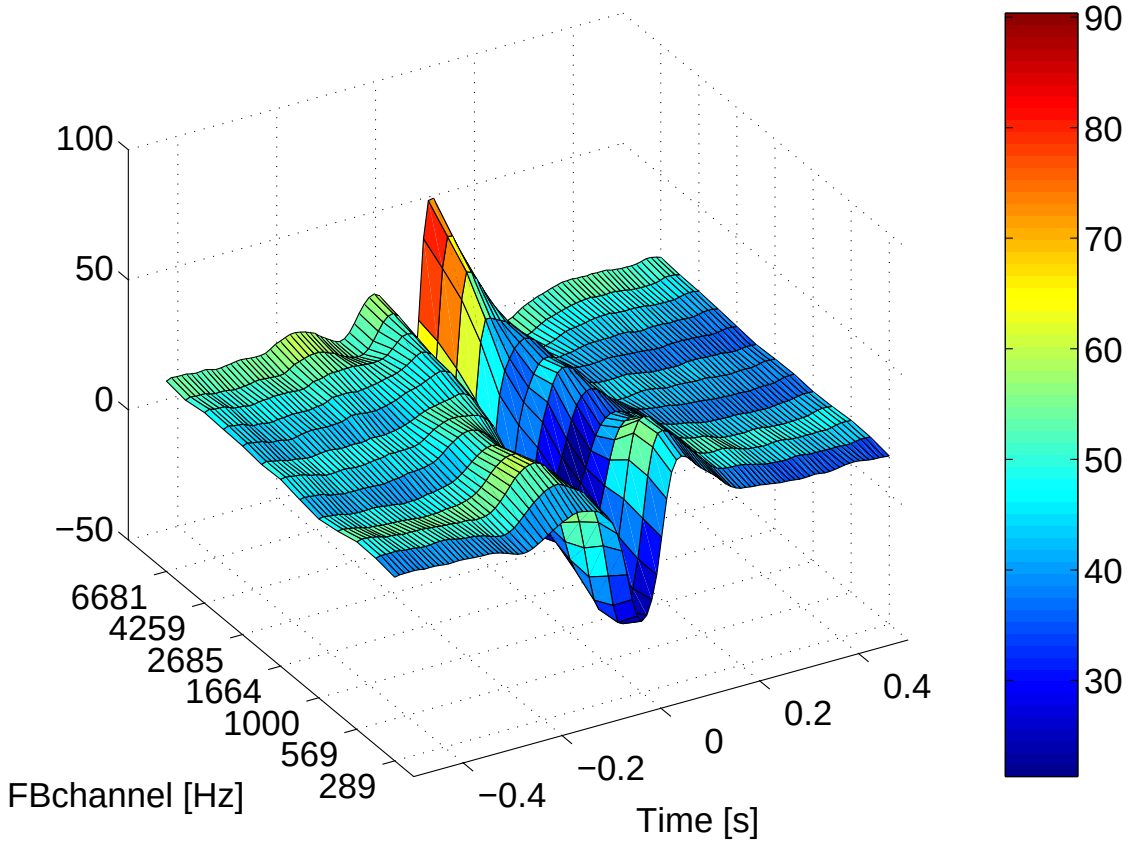


Figure 4.2: PEMO-derived spectro-temporal representation of phoneme /t/. Mean (z-axis) and standard deviation (coloring) in model units, calculated from 6770 instances of this unvoiced stop consonant.

reflect the larger width of articulation bands compared to critical bands (Allen, 1994). Overall, 13 channels are used in the frequency range between 200Hz and 8kHz. The feature vectors are then derived by downsampling the model output to $f_s = 100$Hz. The time signals are processed in context and then segmented using the labels provided with the PhonDat corpus. For each instance a time segment of 101 feature vectors is kept for further analysis, the phoneme center being at the 51th feature vector. Therefore each example is analyzed in about 1s of context.

4.4.3 Secondary Feature Extraction

For the sigma-pi cell parameter values the following restrictions apply: $1 \leq f_n \leq 13$ and $(S - 50) \leq t_n \leq (50 - S)$. In the experiments described below the sigma-pi cells are integrated over time yielding the secondary feature vector. Temporal integration has been varied from $S = 0$ (no

integration) to $S = 20$ (410ms integration). S is kept constant for all features of one individual experiment.

4.4.4 Fisher Score

All possible features are analyzed with respect to their discriminative ability by calculating the Fisher ratio F_i for each individual feature i. The Fisher score is defined as the ratio of between-class variance σ_{inter} over within-class variance σ_{intra}:

$$F_i = \frac{\sigma_{inter}}{\sigma_{intra}} = \frac{\frac{1}{K-1}\sum_{k=1}^{K}(\mu_k - \mu)^2}{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k - K}\sum_{j=1}^{n_k}(x_j - \mu_k)^2} \tag{4.2}$$

with K as the number of classes, N as the total number of examples, n_k and μ_k the number of examples and the mean value for class k, respectively, and x_j as a single observation.

This method is normally used in analysis of variance (ANOVA) if there are more than two groups in a statistical test of significance and in linear discriminant analysis to obtain the basis vectors of the target space. In contrast, in this approach it is assumed that the secondary features already represent a sufficiently good feature space. Therefore, in these experiments a ranking is carried out rather than linear transformation.

4.4.5 Classification and Feature Selection

The final classification unit is kept as simple as possible by using a linear perceptron. The optimal (in RMS sense) weight matrix W is found analytically by calculating the pseudoinverse $X^+ = X^T(XX^T)^{-1}$ of the secondary feature matrix X. The main learning part is the optimal choice of the sigma-pi cell parameters as only a sub-set of all possible cells may be used.

The substitution rule after Gramß (1991, 1992) is used to automatically find an optimal set of secondary features . This algorithm iteratively replaces the least relevant feature by a randomly chosen new one. The relevancy is measured by means of overall RMS error of the target vector Y with and without each feature. In the experiments typically 500 iterations

are performed per training. For each parameter set, five to 20 training runs are carried out with randomly chosen features to start with.

4.5 Evaluation

4.5.1 Feasibility

As a first prerequisite it is to investigate whether short phonemes are distinguishable on the basis of PEMO (primary) features. In Fig. 4.1 and 4.2 the average internal representations (model outputs) of the stop-consonants /d/ and /t/ are shown. The phoneme center is in all cases situated at 0ms. The z-axis represents the mean value for a given 10ms long time segment and filterbank channel, while the color denotes the standard deviation over all instances at this position. It is often argued that the long time constants of the adaption loops in PEMO prevent it from capturing features of e.g. short closures. Obtaining these very different average representations for /d/ and /t/ suggests differently. Other phonemes should be even easier to distinguish due to their longer duration and/or more signal energy.

4.5.2 Parameter Optimization

To check whether this holds in recognition experiments, the FFNN recognizer is trained by substitution rule on a number of different sub-sets of the corpus, each containing only a small number of phoneme classes. In Tab. 4.1 the obtained recognition scores for the test data are shown. Obviously, classification of vowels or diphthongs is much more reliable than classification of consonants. Still, phonemes of different classes can be distinguished quite well as can be seen in the case of a mixed group of phoneme classes. It should be emphasized at this stage that the parameters of the whole system have not been optimized much so far.

The optimal values for temporal integration and the number of secondary features do not vary much over the sub-sets. With more than 30 secondary features the recognition rate does not increase significantly. For $S = 0$, the recognition scores in all cases are significantly worse. It may be concluded that some temporal integration is necessary. This most likely reflects the temporal variation of articulation.

Table 4.1: Recognition scores on test data and the corresponding optimal integration time. Standard deviation over successive runs are denoted in brackets. Results are shown for 30 secondary features.

	classes	rate [%]	S
short vowels	/U/, /a/, /I/	69.3 (4.9)	10
diphthongs	/aI/, /OY/, /aU/	75.8 (6.2)	10
voiced stops	/b/, /d/, /g/	44.4 (3.5)	10
unvoiced stops	/p/, /t/, /k/	46.8 (4.4)	10
mixed	/E/, /AI/, /f/, /n/, /t/	62.7 (5.8)	3

4.5.3 Feature Analysis

To analyze the features which are most relevant for a given task, two different types of experiments may be carried out:

1. Examining the set of 'optimal' secondary features found in the parameter optimization experiment described in 4.5.2.

2. Ranking all possible features by means of the Fisher ratio and analyzing, e.g., the 100 features with the highest Fisher score.

Procedure 1 leads to a set of features which perform well *as a whole* in the given classification task, while using the procedure 2 yields features which *by itself* have high discriminative value. The difference becomes clear by observing that the feature values of the sigma-pi cells with highest score are highly correlated over all training examples. In contrast to that, the substitution rule tends to eliminate similar features in the same set. Still the two types of experiments yield qualitatively comparable results. In this paper, only results from the Fisher ranking experiments are shown.

In Tab. 4.2 the results from the Fisher ranking experiment are shown. For the sub-sets of vowels, diphthongs and the mixed sub-set a maximum fisher score of one or higher is obtained. In addition, the classification performance (on the test set) and the high correlation between the fisher scores on training and test set indicate that the types of sigma-pi cells used in the experiments are suitable as secondary features for classification of vowel-like phonemes and for distinguishing between phonemes of different types.

Another important part of the analysis is to examine the position of the two windows forming the sigma-pi cells. In Fig. 4.3 and 4.4 histograms are

Table 4.2: Results of the Fisher score ranking: the highest value F_{max} of the Fisher scores F_i of all features, correlation coefficient c between Fisher scores on training and test set and recognition performance on the test set by the best N features are listed. In this experiment, temporal integration is carried out with $S = 3$.

		F_{max}		rate [%] for $N =$			
	classes	train	c	2	10	30	100
short vowels	/U/,/a/,/I/	1.92	0.92	37	54	60	82
diphthongs	/aI/,/OY/,/aU/	1.00	0.73	51	63	67	85
voiced stops	/b/,/d/,/g/	0.17	0.25	36	39	48	57
unvoiced stops	/p/,/t/,/k/	0.39	0.44	41	47	61	63
mixed	/E/,/AI/,/f/, /n/, /t/	1.33	0.88	35	44	57	61

shown, counting the number of windows over (absolute) frequency channel and position in time relative to the center of the phonemes. In Fig. 4.3 the 500 sigma-pi cells with the highest Fisher score for three vowels /U/, /a/ and /I/ are taken into account. The highest counts are found for medium frequencies and short temporal distance to the phoneme center, meaning that sigma-pi cells with windows in that region have a much higher Fisher score (i.e. discriminative value) than other sigma-pi cells. Windows near the center of the phonemes (not further than 100 or in some cases 50ms away) are almost exclusively found for the other sub-sets, too (see Tab.4.3). Spectro-temporal patterns further away from the phoneme center might still carry additional information and when looking at the sigma-pi cells derived from the substitution rule experiment (4.5.2), a number of windows are found further away from the phoneme center.

Table 4.3: Statistics of the 100 features with highest Fisher score. Time values are given in ms and frequency values in 2 ERB. Mean and standard deviation (brackets) are shown.

	classes	$\lvert t_2 - t_1 \rvert$	$\lvert f_2 - f_1 \rvert$	t_n	f_n
short vowels	/U/,/a/,/I/	1.7(2.5)	2.8(1.8)	-0.4(2.5)	5.9(2.1)
diphthongs	/aI/,/OY/,/aU/	3.7(2.4)	2.4(1.3)	0.7(3.2)	6.5(1.8)
voiced stops	/b/,/d/,/g/	6.1(2.7)	3.0(2.4)	-0.5(5.3)	8.9(3.5)
unvoiced stops	/p/,/t/,/k/	1.4(1.5)	3.9(2.1)	3.8(2.2)	8.6(2.5)
mixed	/E/,/AI/,/f/,/n/,/t/	2.3(1.9)	3.8(1.9)	-0.4(2.0)	6.1(2.3)

Comparing the window positions for the vowel sub-set to those for the diphthong sub-set (see Tab. 4.3), it is notable that the time difference between the windows tends to be larger for diphthongs, while the frequency difference tends to be smaller. This reflects the difference between de-

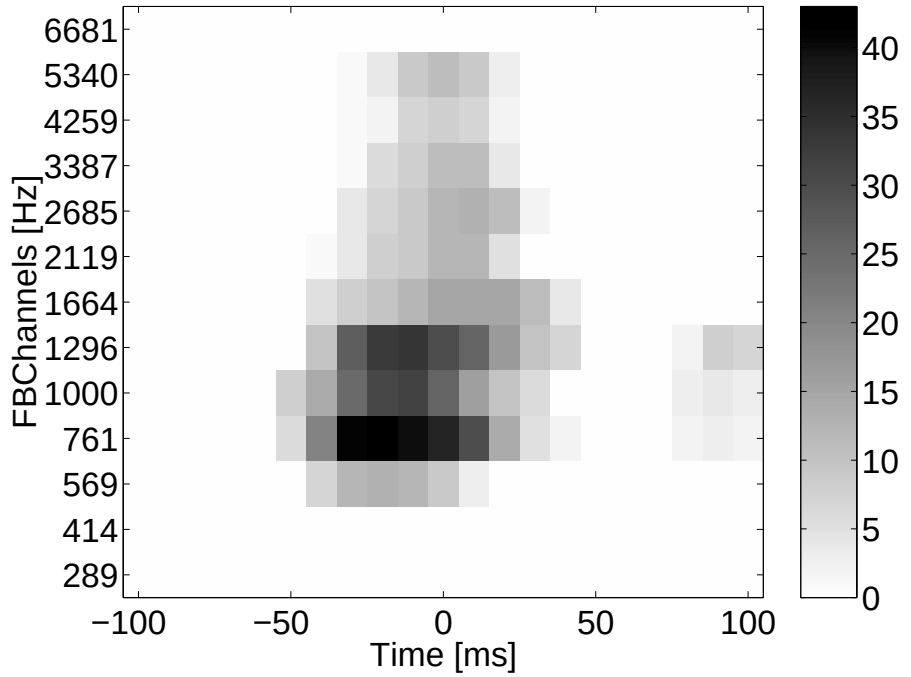


Figure 4.3: Histogram of 500 features with highest Fisher Score for short vowels /U/, /a/, /I/. Window positions relative to center of phoneme are shown, neglecting temporal integration.

tecting simultaneous formants (as in vowels) to formant transitions (as in diphthongs).

In Fig. 4.4 the same histogram is shown for the unvoiced plosives /p/, /t/ and /k/. At least two observations are remarkable: all the windows are at or after the phoneme center and higher frequencies are more important than for vowels or diphthongs. The latter is also observed for voiced stops (see Tab. 4.3) and is probably reflecting the larger amount of energy in high frequency channels. The former might indicate a classification of stops due to their coarticulatory influence on the following phoneme, although a similar effect cannot be found for the sub-set of voiced stops. Alternatively, this might be due to an artifact of PEMO primary feature extraction. In terms of relative window position even higher frequency differences than for vowels can be observed for all stops (up to 10 channels, i.e. almost the whole frequency range).

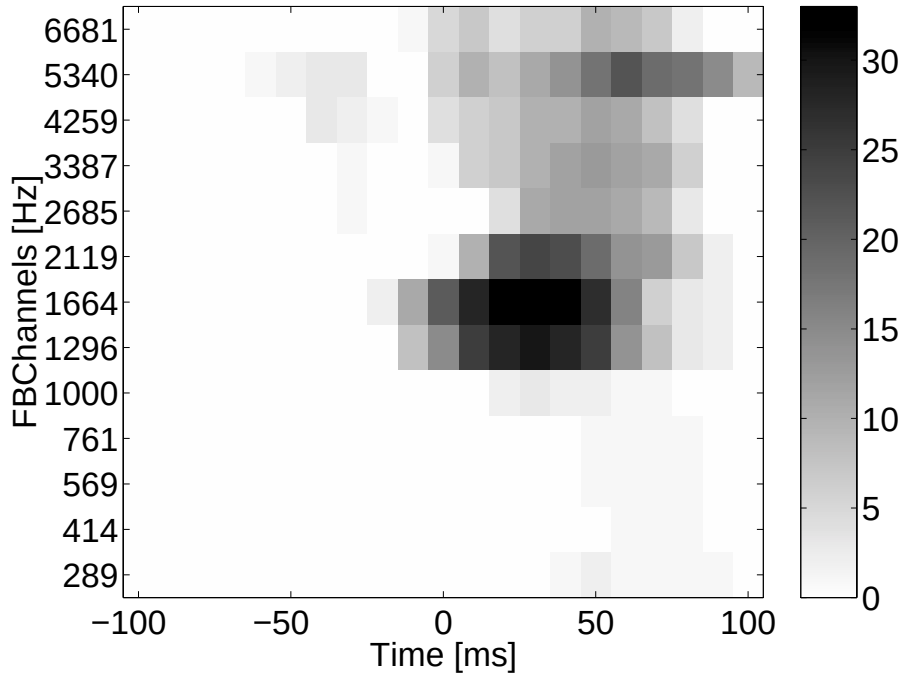


Figure 4.4: Histogram of 500 features with highest Fisher Score for unvoiced plosives /p/, /t/, /k/. Window positions relative to center of phoneme are shown, neglecting temporal integration.

4.6 Discussion

The recognition scores on individual sub-sets of the PhonDat corpus suggest some space for improvements. Still, the sigma-pi cell approach seems to be worthwhile not only in word recognition experiments but also for phonemes. Earlier, yet unpublished studies indicate that the recognition rate may be increased significantly by introducing certain changes (e.g. different lowpass time constants) to the perception model. Furthermore PEMO feature extraction and sigma-pi cells seem to work better on longer segments, e.g. diphones, or context dependent classes. This is for example indicated by the good performance of the system on diphthongs. As presented, an optimized set of sigma-pi cells is easy to analyze and further experiments will be carried out in the future to compare these sets of sigma-pi cells to psychoacoustical and physiological findings.

The author thanks Volker Hohmann and Birger Kollmeier for their support and many fruitful discussions.

chapter 5

Sub-band Signal-to-noise-ratio Estimation Using Auditory Feature Processing [a]

Contents

5.1 Introduction 74
5.2 Feature Extraction 77
5.3 Classifier 84
5.4 Experiments 87
5.5 Discussion 100

Abstract

In this paper a new approach is presented for estimating the long-term speech-to-noise ratio (SNR) in individual frequency bands that is based on methods known from automatic speech recognition (ASR). It uses a model of auditory perception as front end, physiologically and psychoacoustically motivated sigma-pi cells as secondary features, and a linear or non-linear neural network as classifier. A non-linear neural network back end is capable of estimating the SNR in time segments of 1s with a root-mean-square error of 5.68dB on unknown test material. This performance is obtained on a large set of natural types of noise, containing non-stationary signals and alarm sounds. However, the SNR estimation works best for more stationary types of noise. The individual components of the estimation algorithms are examined with respect to their importance for the estimation accuracy. The algorithm presented in this paper yields similar or better results with

[a] A modified version of this chapter was published in *Speech Communication* (39) **1–2**, pp.47–64 (2003) by Michael Kleinschmidt and Volker Hohmann.

comparable computational effort relative to other methods known from the literature for short-term SNR estimation. The new approach is purely based on slow spectro-temporal modulations and is therefore a valuable contribution to both, digital hearing aids and ASR systems.

Zusammenfassung

In dieser Arbeit wird ein neuer Ansatz zur Schätzung des langzeit Sprach-zu-Rausch Verhältnisses (SNR) in einzelnen Frequenzbändern vorgestellt, welcher auf Methoden basiert, die aus der automatischen Spracherkennung bekannt sind. Dabei werden ein Modell der auditorischen Wahrnehmung zur Merkmalsextraktion, physiologisch und psychoakustisch motivierte Sigma-Pi Zellen als sekundäre Merkmale und ein lineares oder nicht-lineares neuronales Netzwerk als Klassifikator verwendet. Ein nichtlineares neuronales Netzwerk kann das SNR in 1 s langen Zeitabschnitten unbekannter Signale mit einem mittleren quadratischen Fehler von 5.68 dB schätzen. Dieser Wert wird auf einem großen Satz natürlicher Störgeräusche erzielt, welcher auch instationäre und Alarmsignale beinhaltet. Dennoch funktioniert die Schätzung am besten für eher stationäre Störgeräusche. Die einzelnen Komponenten des Algorithmus werden bezüglich ihrer Wichtigkeit für den Schätzvorgang untersucht. Der hier vorgestellte Algorithmus erreicht vergleichbare oder bessere Ergebnisse bei ähnlichem Rechenaufwand wie andere Verfahren zur kurzzeit Schätzung des SNR, die aus der Literatur bekannt sind. Der neue Ansatz basiert ausschliesslich auf langsamen, spektro-temporalen Modulationen und ist daher eine wertvolle Ergänzung für digitale Hörgeräte und Spracherkennungssysteme.

5.1 Introduction

The use of digital signal processing in hearing aids offers new, effective means for the rehabilitation of hearing impairment. Apart from higher acoustic signal quality, digital hearing-aids allow for the implementation of specific speech processing strategies, which have no counterpart in the analog domain. For many of these strategies, an estimate of the speech-to-noise ratio (SNR) of the acoustical environment is desirable. Specifically, the following applications are considered: First, many complex algorithms for speech processing are optimized for specific acoustic environments and

might fail in situations that do not meet the specific assumptions. This holds especially for noise reduction schemes that might introduce processing artifacts in quiet environments or in situations where the SNR is below a certain threshold. SNR estimates could be useful to automate the activation or deactivation of processing stages in theses cases. Second, the optimum choice for parameters of compression algorithms and amplification might depend on the SNR, so that an adaptation of the parameter set depending on SNR estimates could help improving the benefit. Third, SNR estimates could directly be used as control parameter for noise reduction schemes (e.g., Wiener filtering or spectral subtraction).

In this paper, feature extraction techniques known from automatic speech recognition (ASR) are applied to the problem of SNR estimation for stationary and non-stationary interfering noise signals. We focus especially on sub-band SNR estimation from a single channel input signal with a temporal resolution from 0.3s to 5s. The temporal resolution considered here is not sufficient for noise reduction by *direct* filtering of the envelope, instead a combination with other faster envelope filtering methods (e.g. on overlap add basis) is required. However, the technique could be useful for the first and second type of applications mentioned above, as well as for improving spectral subtraction algorithms by adding further information about the noise level and its variability. Whereas in the classical spectral subtraction algorithms detection of speech pauses is necessary to estimate the noise spectrum, unconstrained SNR estimation without this necessity is pursued here. In this way, SNR estimation can be extended to non-stationary noise signals that are slowly varying in time.

Different algorithms for unconstrained SNR estimation are known from the literature. Martin (1993) developed an algorithm for SNR estimation for stationary and non-stationary noise signals that is based on the observation of minima of the short time power within time frames of about 0.6s. The algorithm has been applied to sub-band SNR estimation and subsequent spectral subtraction. Hirsch (1993) proposed a method for sub-band SNR estimation based on the observation of histograms of the spectral magnitude within time frames of 250ms to 2s. The system has been tested with car noise, computer room noise and white noise with different bandwidths and was used for speech enhancement. Good results were obtained using time frames of 500ms. This technique was adopted by Avendano et al. (1996) and applied to speech enhancement. Pre-computed FIR filters were selected in each sub-band depending on the respective estimated sub-band SNR in order to increase the SNR. Hirsch and Ehrlicher (1995) extended

the approach by Hirsch (1993) by introducing an adaptive threshold based on the temporal average of the spectral magnitude in each sub-band. The noise level in each sub-band was estimated from the spectral magnitude values below this threshold, either by observing the minima or by observing the histogram of these values. The method was used as noise level estimator in a spectral subtraction noise reduction technique and was tested in automatic speech recognition (ASR) of noisy speech signals. (Bourlard et al., 1996a) adopted the method of observing histograms of spectral magnitudes within sub-bands and assumed that the underlying density distribution has two maxima from speech and noise activity, respectively. The maxima give an estimate of the noise and the speech level and are extracted by energy clustering from the histograms in this approach.

Dupont and Ris (1999) compared different methods for SNR estimation: the histogram method by Hirsch (1993), the weighted average method by Hirsch and Ehrlicher (1995), the energy clustering method by Bourlard et al. (1996a) and a low energy envelope tracking method similar to Martin (1993) that uses the lowest n (about 10) out of M (about 50) spectral energy values for noise level estimation. The envelope follower uses a critical band analysis with a time resolution of 12.5ms, i.e., 50 spectral values cover a duration of 625ms. Dupont and Ris (1999) combined these methods with a narrow-band frequency analysis that allows for estimating the noise levels from the valleys in between the harmonics of voiced speech segments. The system has been tested with amplitude modulated white noise, as well as car and helicopter noise and was applied to ASR. Dupont and Ris (2001) extended this work and supplied data on the comparison of the methods mentioned above and on the application of the methods to ASR. The evaluation was done using different types of noise, especially modulated Gaussian noise, car noise, factory noise and helicopter noise. Time windows were in the range of 250-750ms.

Tchorz and Kollmeier (1999b, 2001) introduced a method for broad band and sub-band SNR estimation that is based on amplitude modulation spectrograms[b] (AMS) for 32ms long time segments. The resulting representation clearly exhibits characteristics related to the harmonic and formant structure of speech (if present). On that basis, a multi-layer perceptron neural network was used as a classifier. Tchorz et al. (2001) applied this AMS-based SNR estimator to noise reduction for ASR applications. The

[b]The amplitude modulation spectrogram as proposed by Kollmeier and Koch (1994) has a frequency and a modulation frequency axis. It is not to be confused with the modulation spectrogram (MSG), which basically is a modulation-filtered spectrogram as described in Section 5.2

estimator by Tchorz et al. demands much more computational effort than most of the methods cited above. Real-time implementation seems not possible yet, but the system can be regarded as a reference system.

This paper extends the methods cited above by applying auditory model based speech feature extraction and neural network classifiers to SNR estimation, that have been shown to increase the robustness of ASR for noisy speech signals (Chapter 3). It is assumed that these robust methods might increase the accuracy of SNR estimates as well, while maintaining a relatively low computational effort, as compared to the system by Tchorz et al.. The approach presented in this paper solemnly exploits low-frequency modulation characteristics of speech in the spectro-temporal domain as no information about harmonicity or high resolution amplitude statistics is present after the primary feature extraction process. This auditory feature based approach only requires low-resolution spectro-temporal patterns.

5.2 Feature Extraction

5.2.1 Material

The speech material is taken from the PhonDat speech database (Kohler et al., 1994). It consists of read German sentences recorded from over 200 native speakers of different dialects. The corpus is divided to form 36 minutes of training and 54 minutes of test material. Training and test part are disjoint with respect to the speakers.

The noise material consists of 41 types of noise in the training set and 54 types of noise in the test set. The types of noise are of variable length and span a range from singing bird to thunderstorm and from alarm clock to machinery and street traffic, all of which are 'natural' sounds and may be encountered in everyday life. Although the noise material consists to the larger part of rather stationary (over a time span of about 1s) types of noise, non-stationary sounds like construction site noise, alarm sounds and babbling speech are also present. Training and test part do not contain the same noise segments. Still, training and test part do contain noise from the same category (e.g. car noise). For further analysis of the SNR classification performance, the noise data has been labeled manually by the authors into a number of categories (see Table 5.1). The category 'stat' describing *rather* stationary noise includes a large variety of different sounds

(for example engine or machinery noise) which in most cases show larger envelope fluctuation than e.g. Gaussian noise. The criterion for labeling a sound as stationary noise was a constant background and no dominating impulsive elements. The categories 'alarm' and 'babble' includes types of noise, which do contain alarm sounds and speech fragments, but may as well contain additional stationary or non-stationary types of noise.

Table 5.1: Frequency of different noise types in the training set (2160 segments of 1s length) and test set (3240 segments). The classification was performed manually by the authors.

category	training		test		description	example
stat	57.2%	(1236s)	80.7%	(2616s)	rather stationary over 1s	car engine
instat	11.8%	(254s)	10.3%	(334s)	rather instationary	constr. site
music	7.9%	(170s)	1.2%	(40s)	containing music	church bells
alarm	3.7%	(80s)	4.0%	(130s)	containing alarm sounds	ambulance
babble	19.4%	(420s)	3.7%	(120s)	containing speech	playground

It is worth noting that the raw speech and noise material that forms the data set for this study is the same as used by Tchorz and Kollmeier (2001). Also the way of dividing into training and test set is left unchanged to allow for a comparison of the results.

5.2.2 Primary Feature Extraction

For sub-band SNR classification the time signal has to be processed and converted into a spectro-temporal representation. The system presented here is based on the perception model (PEMO) after Dau et al. (1996a) which is a psychoacoustical model of the auditory periphery. Other methods of primary feature extraction are examined as possible alternatives or extensions and are described below.

Perception Model

The perception model (PEMO) had been originally developed by Dau et al. (1996a) for quantitatively simulating psychoacoustical experiments, such as temporal and spectral masking. It was adapted for ASR applications and successfully applied as a robust front end in isolated word recognition experiments (Tchorz and Kollmeier, 1999a, Chapter 2) and also serves as a basis for objective speech quality measure (Hansen and Kollmeier, 2000). The ASR-adapted model is sketched in Figure 2.1 on page 25 (Chapter 2).

The first processing stage is a pre-emphasis of the input signal with a first-order high pass filter[c]. This flattens the typical spectral tilt of speech signals and reflects the transfer function of the outer ear. The pre-emphasized signal is then filtered by a bank of gammatone filters each with a bandpass characteristic derived from spectral masking experiments (Patterson et al., 1987). After gammatone filtering, each frequency channel is halfwave-rectified and first order low pass filtered with a cutoff frequency of 1kHz for envelope extraction, which reflects the limited phase-locking ability of auditory nerve fibers above 1kHz. Amplitude compression is performed in a subsequent processing step. In contrast to conventional bank-of-filters front ends, the amplitude compression of the auditory model is not static (e.g., instantaneously logarithmic) but adaptive, which is realized by five consecutive nonlinear adaptation loops (Püschel, 1988). Each of these loops consists of a divider and an RC low pass filter with an individual time constant of 5, 50, 129, 253 and 500ms. Changes in the input signal like onsets and offsets are emphasized, whereas steady-state portions are compressed. The adaptation loops perform a log-like compression for stationary signals and emphasize onsets and offsets of the envelope. This causes a sparse coding of the input in the spectro-temporal domain. The last processing step of the auditory model is a first order low pass filter with a cutoff frequency of 4Hz. Suppression of very slow envelope fluctuations by the adaptation loops and attenuation of fast fluctuations by the low pass filter results in a band pass characteristic on amplitude modulation with a best frequency of about 4Hz (cf. Figure 2.2, page 27, Chapter 2). The primary feature vectors are then derived by downsampling the model output to $f_s = 100$Hz in each channel. The model, as it is used in this study, differs slightly from the original psychoacoustical model and is motivated by automatic speech recognition experiments (Chapter 3). The model adaption includes an additional pre-emphasis and a modified peripheral filter bank, where filter width and spacing are increased to two ERB[d] in order to reflect the larger width of articulation bands compared to critical bands (Allen, 1994). Overall, nine channels are used with center frequencies of 414, 569, 761, 1000, 1296, 1666, 2119, 2685 and 3387Hz, altogether roughly covering the telephone band.

Dau et al. (1997a) proposed a bank of band pass filters for replacing the single modulation low pass filter in order to quantitatively model amplitude modulation perception. The modulation band pass filters are complex

[c] differentiation with factor of 0.97: $y(n) = x(n) - 0.97 * x(n-1)$

[d] equivalent rectangular bandwidth (Moore and Glasberg, 1983)

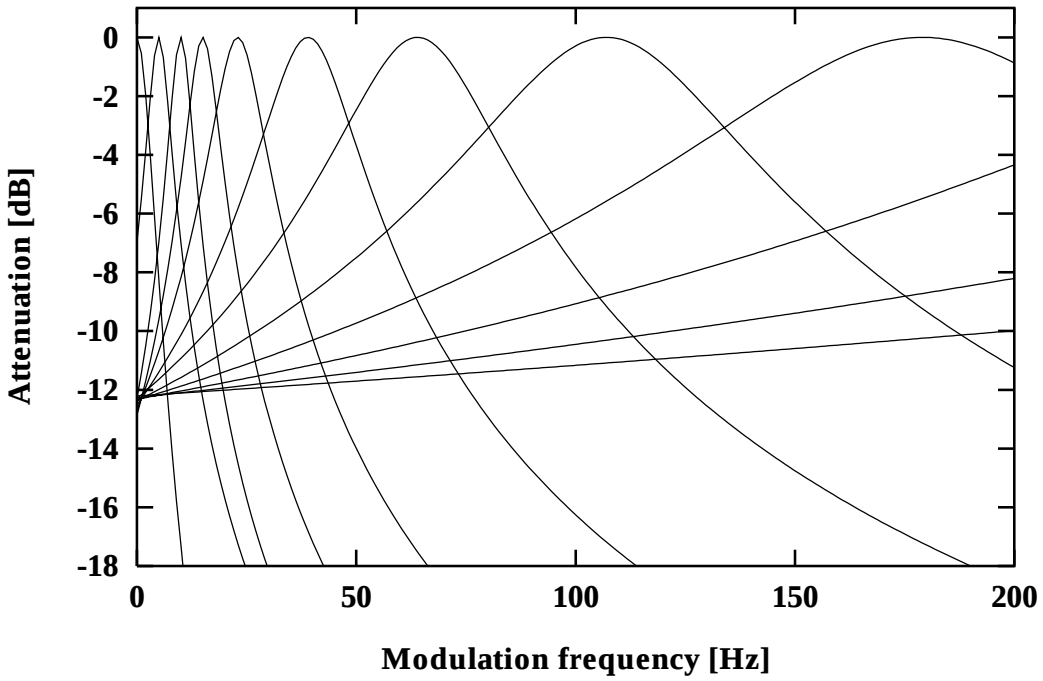


Figure 5.1: Transfer functions of the modulation filters. The modulation filter bank replaces the modulation low pass in PEMO, adding another dimension to the internal representation. Only the range up to 200Hz is plotted. Reprint from Dau et al. (1997a).

first-order IIR bandpass filters. The first modulation filter is a 2.5Hz low pass filter. Below 10Hz there are two more modulation filters with center frequencies of five and 10Hz and a constant bandwidth of five Hz. Between 10 and 1000Hz a logarithmic scaling with a constant Q value of two is applied (see Figure 5.1). In this study only the first three to five (adding 16.6 and 27.7Hz band passes) modulation bands are taken into consideration. For all filters with a center frequency above 10Hz the Hilbert envelope of the modulation filter output is calculated, while in all other cases the real part of the complex output signal is evaluated. For the application in this paper, the modulation filter outputs are downsampled to 100Hz by averaging over 10ms segments.

Modulation Spectrogram

The modulation spectrogram (MSG) has been proposed by Kingsbury et al. (1998) as a robust visual representation of speech which is supposed to change only very little when adding noise or reverberation. In this study a slightly modified version of MSG is used to fit into the overall framework.

In this implementation the peripheral filter bank consists of nine gammatone filters (as for PEMO above) in the range of 300Hz to 4kHz instead of 18 FIR filters of trapezoidal shape. After halfway rectification and 28Hz low pass (1st order IIR instead of a linear-phase FIR filter) the signal in each channel is downsampled to 100Hz (instead of 80Hz) and normalized over the entire signal length. The following complex band pass filter with a passband of 1.25 - 6.72Hz is designed to specifically extract the speech-like modulations of the signal in each frequency band. It is followed by a magnitude to power (in dB) conversion, an overall normalization (maximum set to +15dB) and thresholding (minimum set to -15dB).

Log Energy

As a reference feature extraction, 10ms power average values were derived from the gammatone filter bank output for each of the nine frequency channels, followed by a dB conversion.

5.2.3 Sigma-pi Cells as Secondary Features

Gramß and Strube (1990) proposed sigma-pi cells to be used as secondary features based on critical band spectrograms for isolated word recognition. Sigma-pi cells have later been used in combination with a perception model as front end for isolated word recognition and it could be shown that this combination increases the robustness of ASR systems in additive noise (Chapter 3). It is now to be ascertained whether these psychoacoustically motivated features are suitable for sub-band SNR estimation.

Description of Sigma-pi Cells

Sigma-pi cells are known as second order elements from artificial neural network theory. This term describes certain network units in which the weighted outputs from two or more other units are multiplied before summation over all input values. In this paper sigma-pi cells are always defined on spectro-temporal representations (sequences of primary feature vectors). Each cells consists of two windows of different size with constant distance in time and constant frequency position. The secondary features $s(f_1, f_2, t_0, \Delta t, \Delta f)$ are calculated from the primary feature values $p(t, f)$ as follows:

$$s = \frac{1}{\Delta t \Delta f} \sum_{t} \left[p(t,\ f_1) \cdot \sum_{t'=0}^{\Delta t-1} \sum_{f'=0}^{\Delta f-1} p(t + t_0 + t',\ f_2 + f') \right] \tag{5.1}$$

with f_1 as frequency channel of the small window, f_2 as frequency channel of the lower left corner of the large window, t_0 as the time difference between the windows, and $\Delta t \times \Delta f$ the extension of the large window in time and frequency.

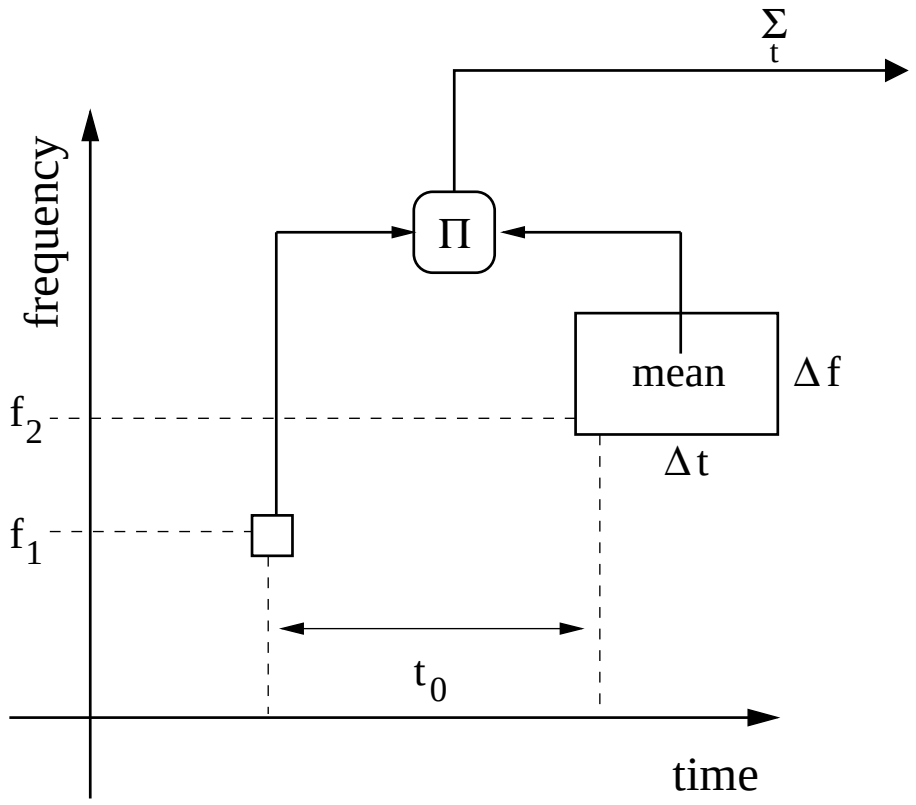


Figure 5.2: Schematic overview of sigma-pi cell calculation from spectro-temporal primary features. For every time step, the small window value is multiplied with the large window mean. The resulting value is integrated over time, resulting in a single secondary feature value.

Note that the small window consists of one element only and that the product of the two windows is integrated over the whole time segment to be classified (see also Figure 5.2). For continuous classification tasks the summation may be replaced by, e.g., a leaky integrator. In the Feature-finding Neural Network (FFNN) proposed by Gramß (1991) the sigma-pi cells form the first processing part of the word classifier. The time invariance problem is solved by giving one of the windows a size larger than one element of the spectro-temporal representation and integrating

over the primary feature vectors of the complete utterance (i.e. word) to be classified. In general, a huge number of different sigma-pi cells are possible within the parameter boundaries (for $f_i \in [1, 9]$, $t_0 \in [-20, 20]$, $\Delta t \in [1, 5]$, and $\Delta f \in [1, 5]$, the number of possible secondary features is over 60.000). An efficient algorithm to automatically obtain a close to optimal set of secondary features is presented in Section 5.3.1.

Why Using Sigma-pi Cells ?

Sigma-pi cells were originally proposed as a part of automatic speech recognition (ASR) systems in order to better capture certain features of speech like formants, formant transitions, fricative onsets and (for larger units) phoneme sequences. A logical "AND" operation is performed by multiplicative combination of the two spectro-temporal windows. This corresponds to the biological counterpart of (cortical) neurons tuned to certain spectro-temporal modulation. It is well known from psychoacoustical threshold experiments that the sensitivity of human listeners to temporal and spectral modulation peaks at around 2-8Hz and 0.25-2 cyc/oct, respectively (Chi et al., 1999). In psychoacoustical reverse correlation experiments, using short segments of semiperiodic white Gaussian noise as stimuli, *early auditory features* of certain spectro-temporal shape were revealed (Kaernbach, 2000). These findings correspond well to physiological measurements of spectro-temporal receptive fields of neurons in the primary auditory cortex (deCharms et al., 1998) which often encompass different unconnected but highly localized parts of the spectrogram.

One may argue whether the resemblance of these basic elements of auditory perception to the spectro-temporal properties of speech is coincidental or not. It is clear that many approaches to robust feature extraction for ASR implicitly take advantage of it. While the approach of calculating temporal delta features in single channels utilizes comb filter effects for selective filtering in the modulation frequency domain, RASTA processing, for example, applies an effective modulation band pass which is comparable to human auditory processing (Hermansky and Morgan, 1994). Although there are efforts to incorporate temporal integration of features over longer time scales into ASR systems (see e.g. Hermansky and Sharma, 1998), true spectro-temporal integration is normally left for the back end recognition system to deal with.

Sigma-pi cells allow for the detection of specific spectro-temporal patterns (consecutive peaks or valleys with a specific distance in any possible direction) on feature level. By using second order features such as sigma-pi cells the secondary feature vectors become even sparser than the already sparse (in the case of the perception model) primary features allowing for easier integration over time and better linear classification. Highly fluctuating noise, alarm sounds and even the superposition of several voices should exhibit different spectro-temporal characteristics compared to a single speech source. This is assumed to be reflected by high values in different sigma-pi secondary features, allowing the classification system to distinguish between foreground speech and other signals. In the case of stationary background noise two effects are expected. On the one hand a constant noise floor reduces the overshoot height of syllable onsets and offsets which dominate the sigma pi cells values. On the other hand the sigma-pi values, which are integrated over time, should show an overall increase because of the additional noise energy in the signal. Considering the nonlinearity of the second order sigma-pi cells, the former effect should dominate, allowing for an estimation of the SNR (cf. Figure 5.7).

5.3 Classifier

In this study, two different neural network classification systems are evaluated. A linear perceptron (LIN) and non-linear multi-layer perceptron (MLP). The linear perceptron is part of the Feature-finding Neural Network (FFNN) framework, which is used to select an optimal set of secondary features. The optimal set is then used as input to the MLP.

5.3.1 Feature-finding Neural Network

Gramß and Strube (1990) proposed to use a linear single-layer perceptron in conjunction with secondary feature extraction. The resulting classification system, called *Feature-finding Neural Network* (FFNN), was applied to isolated word recognition tasks. Their argument was twofold:

1. for a sufficiently high-dimensional feature space (i.e. a large number of secondary features), a linear net yields equal or better classification and generalization results when compared to a non-linear classifier.

2. a linear classifier allows for easy and fast training. The error function is hyper-paraboloid with one minimum.

Given P examples, each represented by a secondary feature vector with M elements, the feature vectors form a $M \times P$ feature matrix $\mathbf{X}$. Given the target matrix $\mathbf{Y}$ ($N \times P$ with N as the number of target values per example), the optimal (in RMS sense) weight matrix $\mathbf{W}$ ($N \times M$) is found analytically by calculating the pseudo-inverse

$$\mathbf{X}^+ = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{5.2}$$

of the secondary feature matrix $\mathbf{X}$. The weight matrix is obtained as

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^+ \tag{5.3}$$

and minimizes the classification error

$$E = |\mathbf{Y} - \mathbf{W}\mathbf{X}|^2. \tag{5.4}$$

An easy-to-train classifier is exactly what is needed to automatically find a sub-set of all possible secondary features, which is optimal or at least well suited for a given classification task. Gramß (1991) proposed a number of training algorithms for the FFNN system, one of which, the *substitution rule*, is used in this study:

i. Choose M secondary features arbitrarily.

ii. Find the optimal weight matrix $\mathbf{W}$ using all M features and the M weight matrices that are obtained by using only $M-1$ features, thereby leaving out every feature once.

iii. Measure the relevance R of each feature i by

$$R_i = E(\text{without feature } i) - E(\text{with all features}) \tag{5.5}$$

iv. Discard the least relevant feature $j = \text{argmin}(R_i)$ from the sub-set and randomly select a new candidate.

v. Repeat from point (ii) until the maximum number of iterations is reached.

vi. Recall the set of secondary features that performed best on the training/validation set and return it as result of the substitution process.

Although the classification is performed by a linear neural network, the whole classification process is highly non-linear due to the second order characteristics of the sigma-pi cells and the restriction to a small sub-set of all possible secondary features. After optimization the set of secondary features may be analyzed with respect to sigma-pi cell parameters and is relatively easy to interpret. The thereby obtained set of secondary features might also be used as input to other, more sophisticated classification systems.

5.3.2 Multi-layer Perceptron

While the linear classifier has advantages in optimizing the set of sigma-pi cells and presumably is sufficient for separating several classes by hyperplanes in multi-dimensional (secondary) feature space, it is, by definition, not capable of non-linear transforming the secondary feature input values to continuous SNR estimates.

To allow for non-linear classification, the linear perceptron is replaced by a multi-layer perceptron (MLP) after obtaining a set of secondary features using the FFNN framework. The MLP consists of two layers (hidden, output). Before feeding into the MLP, the secondary feature and output target values are first normalized (zero mean and unit variance) and the former also decorrelated by principal component analysis (PCA), thereby reducing the number of input neurons, i.e. numerical complexity, significantly. The PCA is carried out on training data. During testing (application) the PCA stage is represented by a simple linear transform, i.e. matrix multiplication, increasing the computational complexity only by a small amount.

Although the performance does not vary much over a reasonable choice of number of hidden and input units, best results were obtained by using the twelve PCA dimensions with the highest Eigenvalues (out of 20 parameter dimensions) and designing the MLP to have 14 hidden units. The two neural networks do not differ much in the number of free parameters. The perceptron (LIN) has 180 weights, while the MLP has 294.

The transfer functions of this feed-forward network were set to *tanh* for the hidden layer neuron and linear in the output layer. Net weights were

calculated by Levenberg-Marquardt Training[e] (Hagan and Menhaj, 1994). 20 iterations of this second-order training method were sufficient.

5.4 Experiments

5.4.1 Setup

Speech and noise material are low pass filtered and downsampled to 8 kHz sampling frequency, high pass filtered[f] and finally added with overall broad band SNR of 0 and 10dB, each for half of the material. The segmental (short-term) sub-band SNR exhibits a wide distribution of values with a maximum at about 5dB (see Figure 5.3). To compensate for possible level-dependency effects, the training and test material was systematically attenuated by between 0 and 34dB.

Before mixing speech and noise, target values for the speech-to-noise ratio (SNR) are calculated in dB for nine sub-bands derived from the ratio of RMS values of speech and noise, respectively, over the whole time segment of, e.g. 1s. The range of SNR values is restricted to the relevant interval of -10 to 20dB as by Tchorz and Kollmeier (2001). The sub-band signals are formed by the output of a gammatone filter bank using the implementation of Hohmann (2002). The nine filters in the frequency range between 300Hz and 4kHz match the peripheral filter bank of the primary feature extraction (see Section 5.2 above). The gammatone filter bank in the given configuration is chosen because it covers the most important frequency range of speech, which is of interest for telecommunication and hearing aid applications. The filter bank divides it into a reasonable number of sub-bands, comparable to what is done in hearing aid compression algorithms. The mixed time signals are processed in three minute long parts and the resulting primary feature vectors are then cut into segments of one second length resulting in 2,160 instances times 9 frequency channels = 19,440 local SNR values in the training set and 3,240 instances times 9 channels = 29,160 target values in the test set.

A set of suitable sigma-pi cells is derived by means of FFNN applying the substitution rule with 1500 iterations. The following constraints are applied to the secondary features: $f_1 \in [1, 9]$, $f_2 \in [1, 9]$, $t_0 \in [-20, 20]$,

[e] from the MATLAB Neural Networks Toolbox

[f] 4th order Butterworth high pass with cutoff at 250Hz

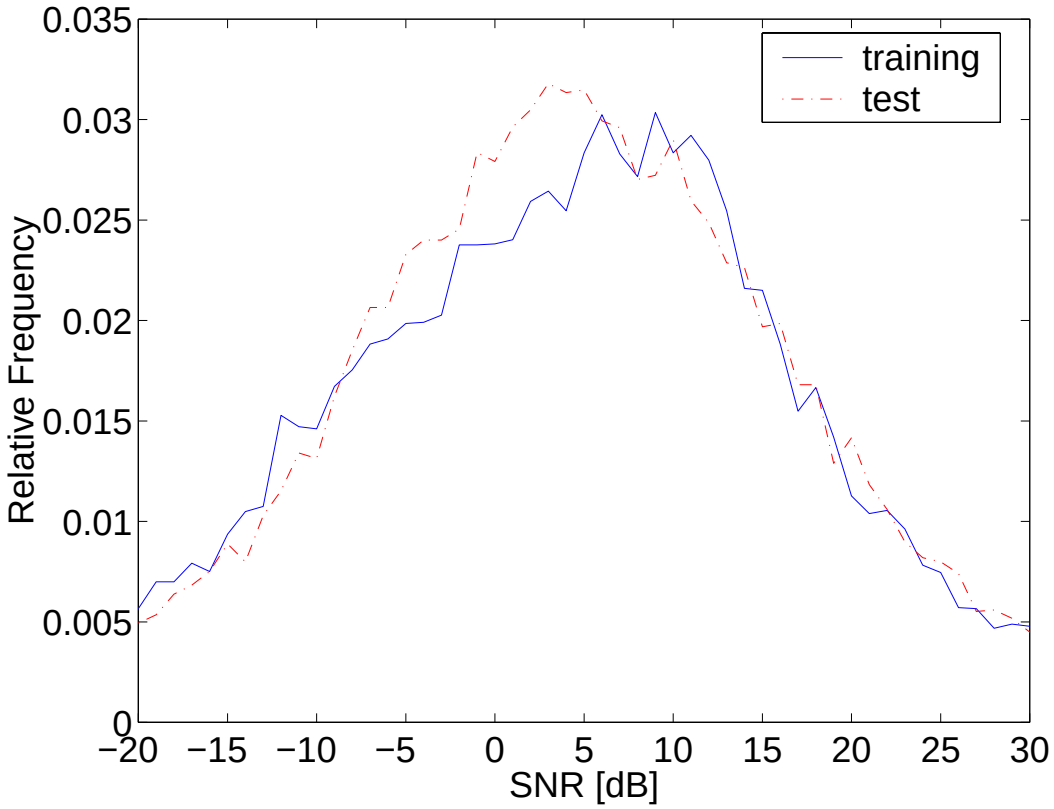


Figure 5.3: Histogram of segmental SNR occurrence in all nine frequency channels for 1s long segments of training and test material. The overall standard deviation of the SNR values in the test set is 10.5dB, ranging from 9.9 to 10.6dB in individual frequency channels.

$\Delta t \in [1, 5]$ and $\Delta f \in [1, 5]$. Finally, the MLP is trained using the already optimized secondary feature set. Best results are obtained with 12 input, 14 hidden and 9 output neurons. For both networks the training material also served as validation data in the training process.

The results are given in the following measures of error to allow for comparison with the literature (x denotes SNR in dB, E the expectation value) :

- estimation error $X = x_{estimated} - x_{true}$
- mean absolute deviation $E(|X|)$
- mean deviation (bias) $E(X)$
- root-mean-square (RMS) error $\sqrt{E(X^2)}$
- statistical error $\sigma = \sqrt{E(X^2) - E(X)^2}$.

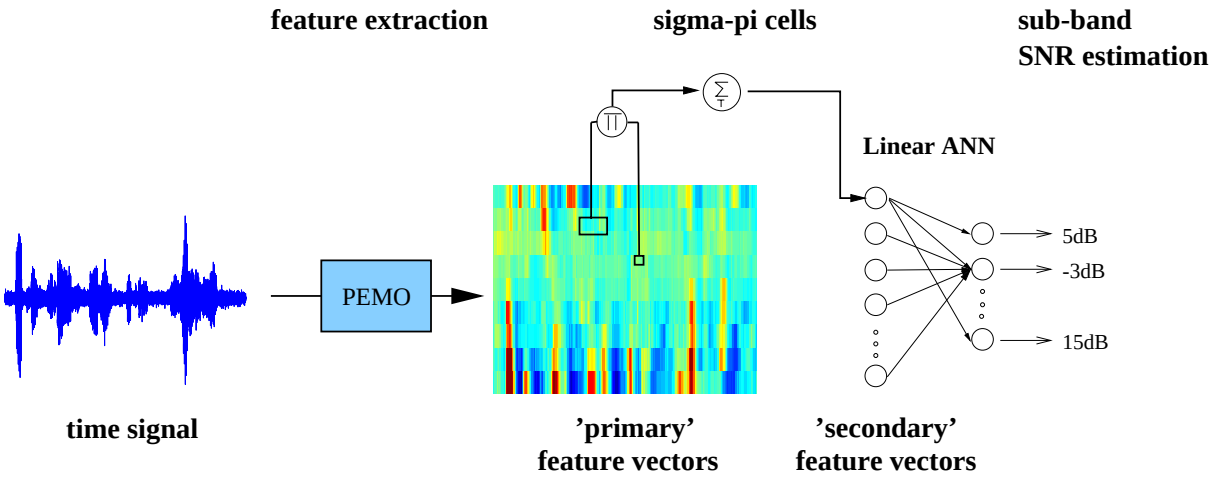


Figure 5.4: Overview of processing steps in the SNR estimator. From a segment of the time signal a sequence of primary feature vectors is extracted by the perception model (PEMO). For each sigma-pi cell one single secondary feature value is derived by integrating over time. The resulting secondary feature vector is linearly transformed into sub-band SNR estimates by the single layer perceptron (LIN). Better performance may be obtained by replacing the linear network with a nonlinear multi-layer perceptron (MLP).

5.4.2 Results

The combination of PEMO, sigma-pi cells and MLP is able to predict the sub-band SNR with an overall RMS error (all types of noise and all bands) of 4.54dB on the training data and 5.68dB on the test data. The performance degrades by about 2dB when the linear neural net is used instead of the MLP (see Section 5.4.3 below).

In Figure 5.5 the median and 10/30/70/90 percentiles of estimation error are shown for 1dB wide bins of true SNR value. For very low SNR values (-5dB and smaller) there is a tendency to overestimate the signal-to-noise ratio, while for very high SNR values (12dB and higher) underestimation can be observed. Two factors are likely to influence this effect. First, the distribution of SNR values peaks at about 5dB (Figure 5.3), so that very low and very high SNR values are underrepresented in the training data and the estimation is biased towards the center of distribution. Second, limiting the measured SNR values to the range from -10 to 20dB should also result in a skewness of error distribution. This is clearly seen from the 10/90 percentiles. A similar effect was observed by Tchorz and Kollmeier (2001) to a much larger extend (up to 6dB) than the systematic bias of maximal 2dB absolute in the current approach.

Looking at individual frequency channels (cf. Figure 5.6), the estimation is performed with similar success in all nine sub-bands and the bias is

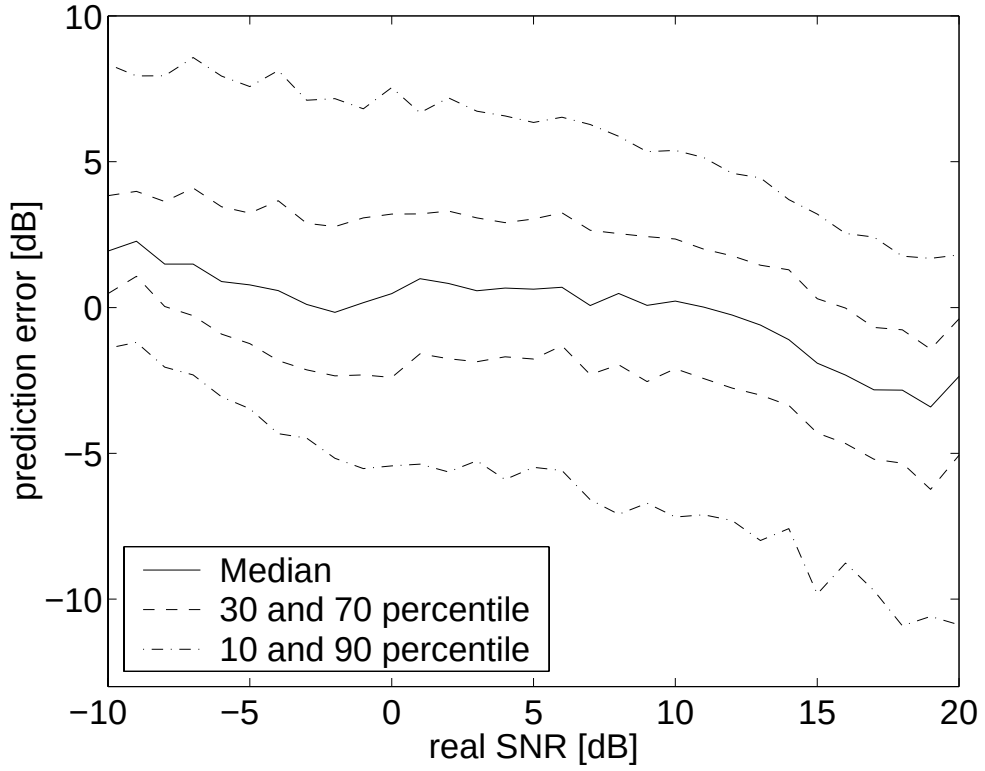


Figure 5.5: Test data set: Percentiles of estimation error distribution for all frequency channels and all noise categories in 1dB bins of real SNR.

close to zero for all bands. For configurations with other front ends, there is a trend towards higher errors in the lowest and the highest filter bank channel. This is probably due to less possibilities in utilizing covariance of neighboring channels or spectro-temporal features for channels at the bottom or top frequency margin.

The performance strongly varies over the five noise categories. The results summarized in Table 5.2 show that the SNR estimation performs best for music, stationary and babble noise, while yielding much higher error values for alarm sounds or non-stationary types of noise. The SNR value are overestimated for the categories 'instat' and 'alarm' by 3.10 and 2.26dB, respectively. For 'babble' the mean absolute deviation of 4.31dB is almost completely explained by a systematic bias of 3.53dB. It is not surprising that the algorithm tends to confuse babble and alarm sounds with the foreground speaker, as the spectro-temporal modulation properties of these sounds are similar to those of speech. Dominating non-stationary components typically contain sharp broad band impulses with a common onset and offset in most frequency channels. The estimation algorithm seems to have more difficulties in this case than for rather stationary types

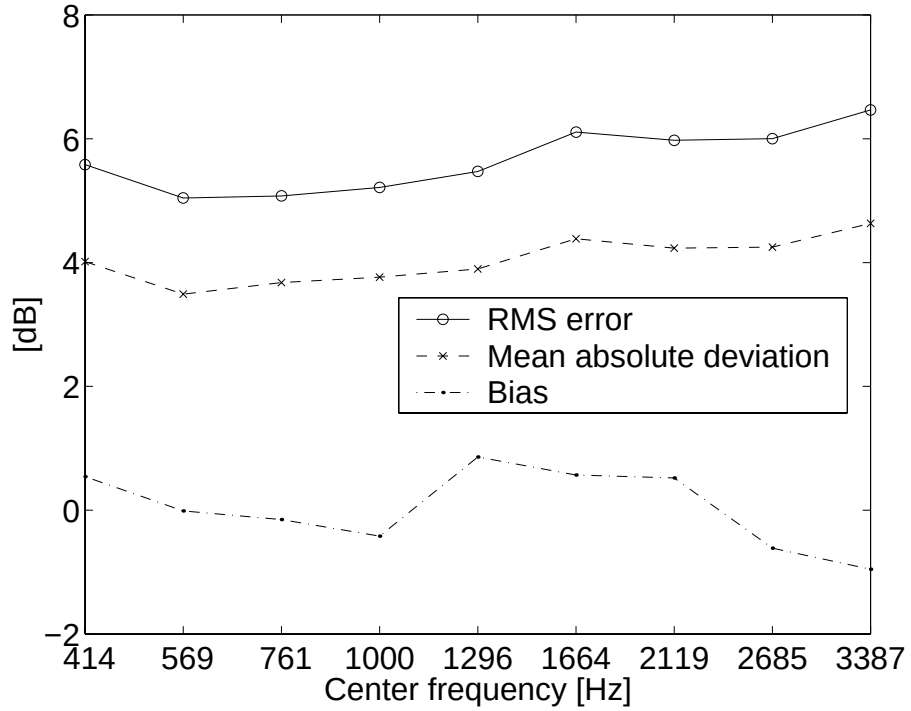


Figure 5.6: Test data set: RMS error, mean absolute deviation and bias for individual frequency channels and all noise categories.

of noise. This could be partly due to the over-representation of stationary noise in the training data. A more detailed analysis is presented in Section 5.4.7 below.

Table 5.2: SNR estimation error (RMS, mean absolute deviation and bias) in dB on test data. The performance is calculated over all frequency channels. Values are given separately for different noise categories. The overall performance ('all') is compared to the results obtained by Tchorz and Kollmeier (1999b) on the same data set.

category	RMS $\sqrt{(E(X^2))}$	deviation $E(\|X\|)$	bias $E(X)$
stat	5.04	3.68	-0.63
instat	8.50	5.93	3.10
music	4.91	3.85	0.20
alarm	8.72	6.15	2.26
babble	5.20	4.31	3.53
all	5.68	4.04	0.04
Tchorz		5.4	

The overall performance is compared to Tchorz and Kollmeier (1999b), who used the same data set. The mean absolute deviation of the perceptual feature based system is about 1.4dB less than the mean absolute

deviation of the AMS-based SNR estimator. However, one should be careful comparing the two algorithms as the time scales are not identical (1s vs. 32ms). Tchorz and Kollmeier (2001) showed that simple low pass filtering of the estimate with a time constant of 1s typically yields an improvement of 1.5 to 2dB in mean absolute deviation and the performance of the algorithm described in this paper slightly decreases for shorter time segments (see Section 5.4.5).

By calculating a noise level estimate from the SNR estimation and the input signal level, the perceptual feature based sytem may, in principle, be compared to the algorithms examined by Dupont and Ris (2001) which are described in the introduction. The performance seems to be comparable for rather stationary types of noise, while for some slowly fluctuating types of noise (Gaussian noise modulated with 0.5 or 1Hz) the perceptual feature based method yields better results. This is easily explained by the main difference between the experimental paradigms used in this paper and by Dupont and Ris (2001), namely the target values, which are calculated on time scales of 1s and 32ms, respectively. Still it may be concluded that the perceptual feature based system yields a reasonable performance on the long-term sub-band SNR estimation task, which is roughly in the range of other approaches.

5.4.3 Primary Feature Dependency

In order to investigate which primary feature extraction method is most suited for SNR estimation in the given framework, PEMO has been replaced by other processing schemes. Overall RMS error values are presented in Table 5.3. The results indicate that PEMO feature extraction yields smallest error values, both with a linear and non-linear neural network classifier. Especially the MSG and the log-energy front end are outperformed by primary features which contain PEMO processing.

Also, it should be noted that adding other primary features to the original PEMO front end does not result in better estimation performance. Both, the combination of PEMO with MSG and the replacement of the modulation low pass by modulation filter bank yields stagnating results at best, partly even an increase in error can be observed. When PEMO with an additional modulation filter bank is used as a front end, the performance on the training data increases for MLP, while the results on the test data

Table 5.3: RMS error in dB for SNR estimation on training and test data with linear (LIN) and non-linear (MLP) neural network using different methods for primary feature extraction. PEMO-ModFB3 and PEMO-ModFB5 denote PEMO with first three and five modulation filters, respectively.

	LIN		MLP	
Primary feature	train	test	train	test
PEMO	6.49	7.33	4.54	5.68
MSG	6.89	8.28	6.07	8.07
LOG-Energy	6.75	7.73	6.19	8.25
PEMO-ModFB3	6.61	7.38	4.28	5.97
PEMO-ModFB5	6.51	7.23	4.19	6.03
PEMO+MSG	6.46	7.23	4.69	6.06

degrade slightly. This indicates a loss of generalization ability with a rising number of elements in the primary feature vectors.

These results confirm findings from ASR applications that the combination of PEMO and sigma-pi cells is especially rewarding in terms of recognition performance. In Figure 5.7 the primary feature vector sequence (PEMO output) is plotted for clean speech and the same signal mixed with machinery noise at 10dB SNR. Although the onset peaks remain clearly visible, their height is reduced and valleys are 'filled' with background activity, resulting in different sigma-pi cell values and thereby allowing a neural network to estimate the SNR.

5.4.4 Secondary Feature Dependency

As a further analysis of the SNR estimation scheme, the number of secondary features has been varied. Starting from a set of 60 sigma-pi cells optimized by the substitution rule, one cell after another is discarded in the order of rising relevance R starting with the least relevant. The results are shown in Figure 5.8 for the linear classifier (LIN) and in Figure 5.9 for the non-linear feed forward network (MLP).

The MLP-based SNR estimator shows no further increase in performance with the number of secondary features exceeding 20. A slightly different behavior is observed for the linear classifier, which gains some benefit from higher number of sigma-pi cells, although this effect is more prominent with the training data. It can be concluded that with a non-linear classifier, 20

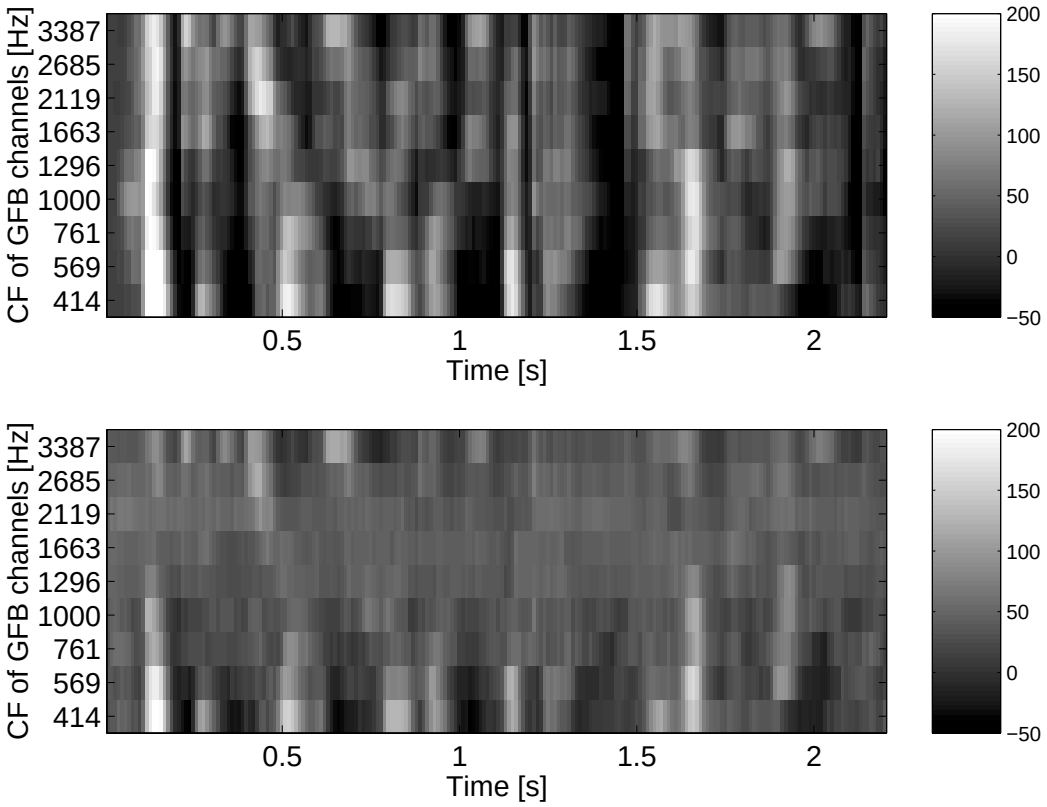


Figure 5.7: Spectro-temporal representation (PEMO primary features) for the German sentences *'Heute ist schönes Frühlingswetter. Die Sonne lacht.'* (It's fine weather today. The sun is shining brightly.) clean (above) and mixed with machinery noise at an SNR of 10dB (below). Note that for equal scaling of the output (gray shading denotes output values in model units) the peaks and valleys are less prominent when background noise is present.

secondary features are sufficient for estimating the SNR in nine frequency channels.

In further experiments it is examined how important different types of sigma-pi cells are for the SNR estimation process. This is done by altering constraints for the parameters. By allowing only temporal features (condition $f_1 = f_2$), spectral integration on the secondary feature level was excluded ('temporal'). With $t_0 = 0$ the focus is on spectral integration ('spectral'). By changing the parameter range to $f_1 \in [1, 9]$, $f_2 \in [1, 9]$, $t_0 \in [-30, 30]$, $\Delta t \in [1, 7]$ and $\Delta f \in [1, 7]$ a larger variety of sigma-pi cells was presented to the optimization algorithm ('large') than in the standard setting.

The results in Table 5.4 indicate that spectro-temporal integration is necessary to obtain full performance with the MLP-based estimator. Restricting the set of possible features to temporal or spectral features only leads to a

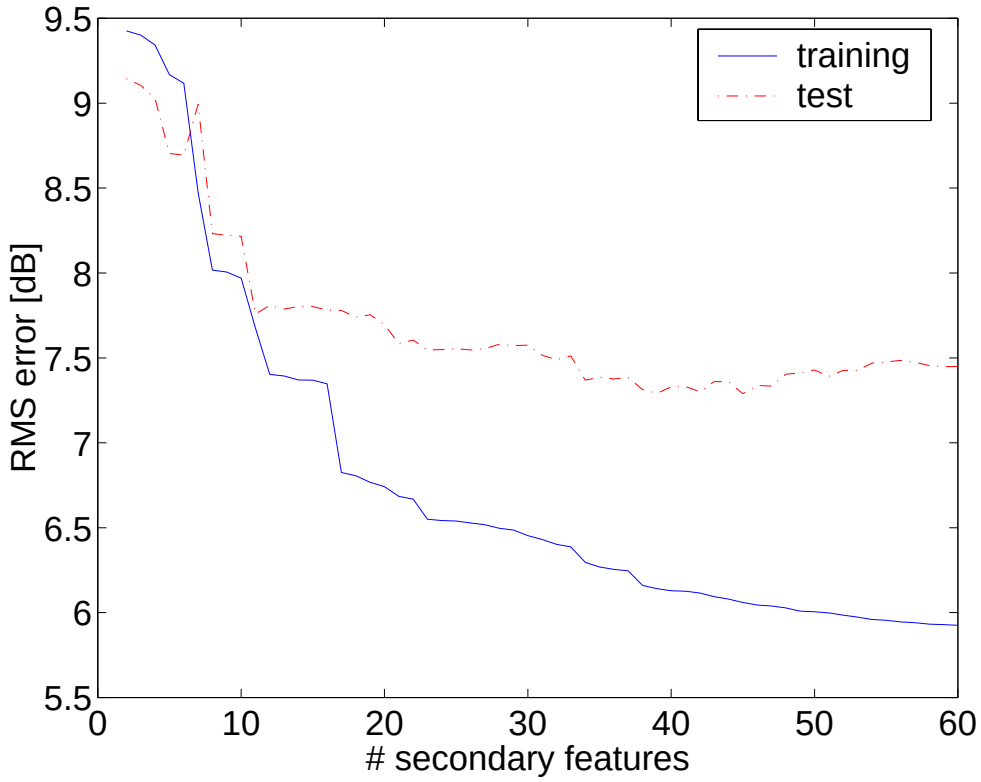


Figure 5.8: RMS classification error on training and test data for linear neural network (LIN) depending on the number of secondary features used as input.

Table 5.4: RMS error in dB for SNR estimation on training and test data with linear (LIN) and non-linear (MLP) neural network depending on the type of sigma-pi cells allowed. The *standard* set is compared to sets of *temporal* features $f_1 = f_2$ and *spectral* features $t_0 = 0$ only, as well as to a *large*r set of possible features (see text).

	LIN		MLP	
feature set	train	test	train	test
standard	6.49	7.33	4.54	5.68
large	6.48	7.12	4.67	5.97
temporal	6.57	7.34	5.12	6.70
spectral	6.46	7.35	5.27	6.74

decrease of about 1dB. This is only observed for MLP. The linear classifier is not at all affected by these constraints yielding almost constant error values. Also, allowing a larger set of possible features does not change the performance significantly.

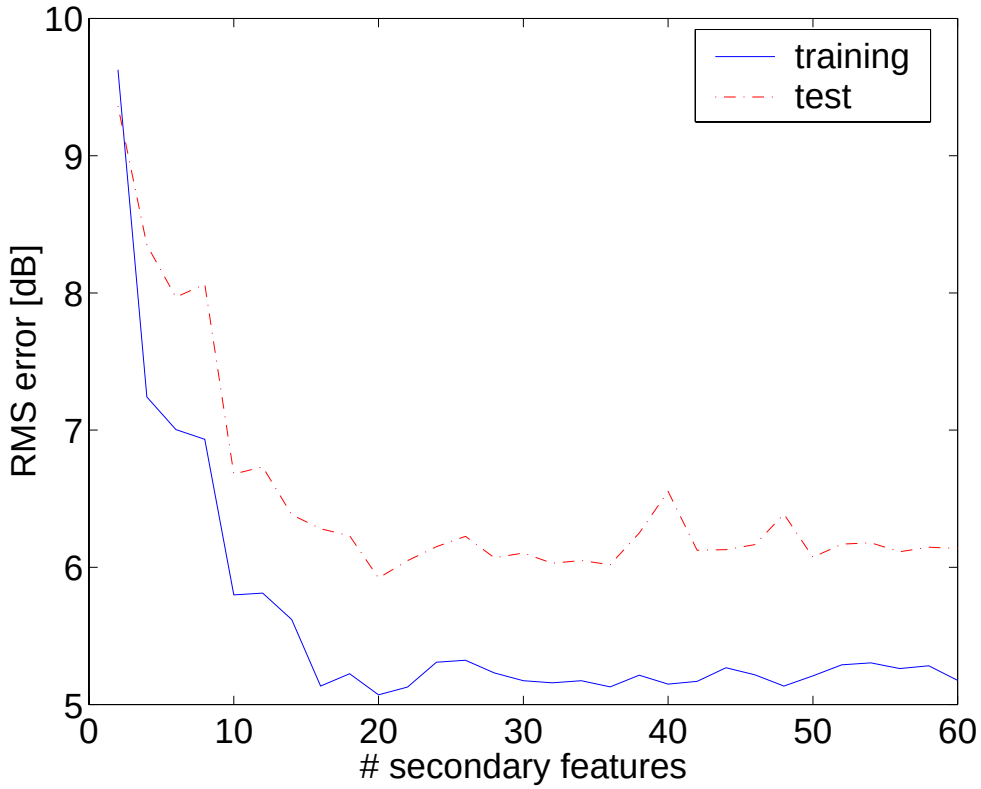


Figure 5.9: RMS classification error on training and test data for multi-layer perceptron (MLP) depending on the number of secondary features used as input.

5.4.5 Segment Length Dependency

So far, all experiments are carried out with a segment length of one second. As for some applications shorter or longer time scales are of interest, the segment lengths is systematically varied between 0.3 and 5s. With a maximal time difference of 200ms between the two windows of the sigma-pi cells and large windows of up to 50ms temporal extension in the standard setup, shorter time segments than 300ms are not feasible. For shorter time segments the choice of sigma-pi cells would have to be restricted further.

The results in Figure 5.10 indicate that for segment lengths of 0.8 to 5s the estimation error is almost constant. The performance degrades towards shorter time segments by up to 1dB RMS error for 300ms length. With shorter time segments the probability of finding specific spectro-temporal patterns, namely the ones the individual sigma-pi cells are tuned to, declines. Therefore an increase in estimation error is expected, even if speech is present over the whole segment duration.

It should be noted that the number of examples (segments) in training and test data varies with the segment length while the absolute duration of the

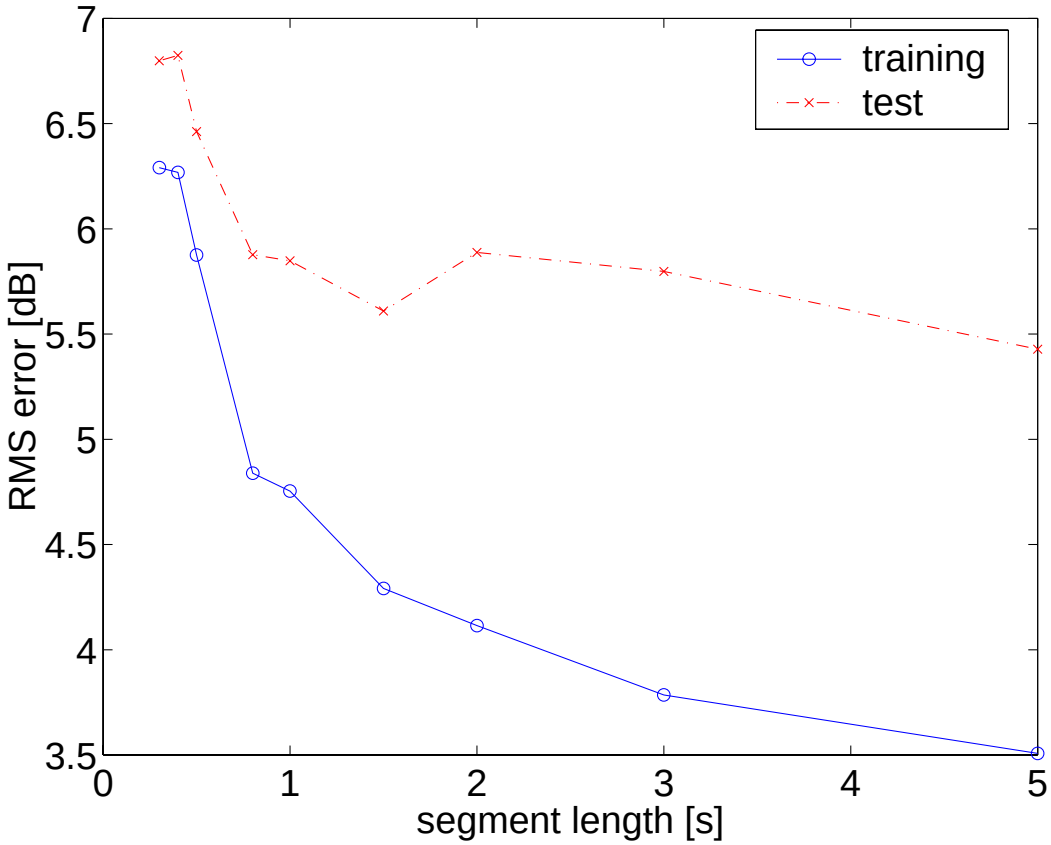


Figure 5.10: RMS classification error on training and test data for MLP depending on the segment length.

training and test material remains constant. With lower number of training examples, i.e longer time segments, a neural network is prone to over-training, resulting in lower error for the training material and stagnated or even degraded performance on the test data.

5.4.6 Computational Effort

Apart from the classification performance, the amount of memory storage and the computational complexity are the most important criteria for the feasibility of an algorithm. This is especially true for a battery driven mobile device, such as a hearing aid or mobile phone. To put the proposed algorithm into perspective, Table 5.5 lists estimates for the order of magnitude of operations required for the perceptual feature based SNR estimator and other algorithms from the literature.

This clearly shows that the proposed sigma-pi/MLP back end of the SNR estimator is comparable or less computationally demanding than most

Table 5.5: Estimated computational requirements in 1,000 operations per second for different front ends (*cursive*) and sub-band SNR or noise level estimation algorithms. The values are given in four categories. ADD contains additions, subtractions and comparisons, MULT multiplications, DIV divisions and FUNC any non-standard operation (logarithm, square-root or sigmoid), that takes longer time to compute ,e.g., by using a table look-up or a Taylor series expansion. The estimates are given in order of magnitudes only, because the exact numbers vary depending on implementation parameters. Feature extraction algorithms are listed in *italic* Font.

algorithm	ADD	MULT	DIV	FUNC
Dupont and Ris (2001):				
short-term FFT power	*100*	*100*	*0*	*10*
Hirsch histograms	10	10	0	0
Weighted average	100	10	1	10
Low E envelope tracking	1,000	100	0	0
energy clustering	10,000	10,000	0	1,000
Tchorz et al. (2001):				
AMS pattern calculation:	*10,000*	*1,000*	*0*	*10*
MLP back end	1,000	1,000	0	10
perceptual feature based:				
PEMO	*1,000*	*1,000*	*100*	*0*
Sigma-Pi / MLP	100	10	0	1

other algorithms. While the PEMO front end exceeds the computational complexity of FFT-based power spectrogram (for short windows of the order of 10ms in length) by one order of magnitude, it requires considerably less computations than the amplitude modulation spectrograms (AMS) used by Tchorz et al. (2001). The primary feature extraction with PEMO is the only part of the perceptual feature based SNR estimator that is computationally expensive. It might be very well possible to replace it with simple contrasted spectrograms (Gramß and Strube, 1990) or less complex PEMO variants with e.g. two instead of five adaptation loops (as examined by Tchorz and Kollmeier, 1999a) - both are ways of saving considerably amounts of computation time.

5.4.7 Importance of Temporal Modulation

Temporal modulation is a very important factor for the SNR estimator proposed in this paper because sigma-pi cells are designed to specifically detect certain spectro-temporal modulations, depending on the window

size and distance in time and frequency. The noise categories differ in their envelope modulation properties which might explain the differences in SNR estimation accuracy found between different types of noise (Section 5.4.2 above).

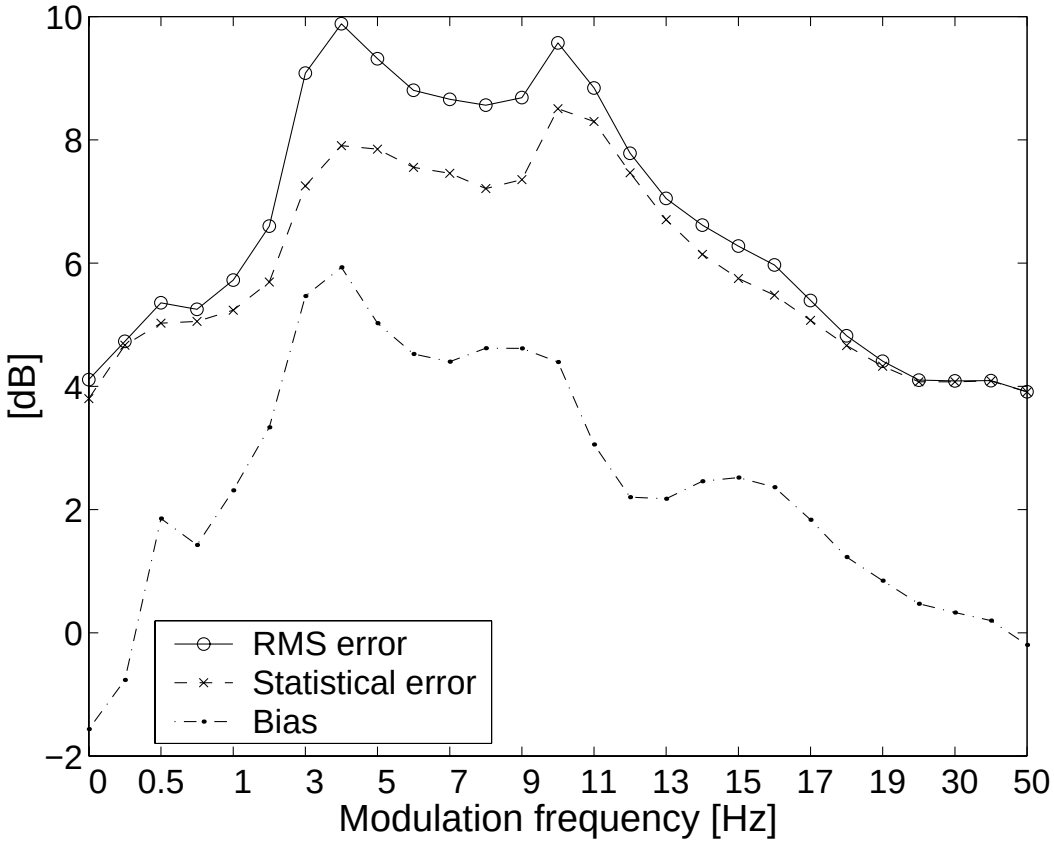


Figure 5.11: RMS error $\sqrt{E(X^2)}$, statistical error σ and bias $E(X)$ for SNR estimation on a three minute long speech signal with Gaussian noise of varying modulation frequency.

In order to investigate the influence of temporal envelope fluctuations systematically, SNR estimation experiments are carried out with Gaussian noise as a carrier and sinusoidal modulations (m=1, full modulation) of various frequencies between 0 and 50Hz. This fully modulated broad band noise signal is a very difficult task for SNR estimation algorithms and quite artificial. Nevertheless, it was chosen here, because it helps analyzing the variability in SNR estimation performance observed for different noise categories.

In Figure 5.11 RMS error values are plotted over modulation frequency f_m. As expected the overall error is lowest for very slow ($f_m \leq 1$Hz) and very fast ($f_m \geq 15$Hz) envelope fluctuations. The error function peaks at around 4Hz and again at about 10Hz. These peaks correspond to the syllable and

phoneme frequency in speech. Assuming that the sigma-pi cells are tuned to certain spectro-temporal patterns of syllabic or diphonic units, it can be explained why speech is hard to distinguish from these broad band modulated noise signal, resulting in a systematic overestimation of the SNR by up to 6dB.

5.5 Discussion

A novel perceptual feature based SNR estimator is introduced that uses spectro-temporal modulations only. The harmonicity of voiced speech is not exploited by this approach as the primary feature vector sequence forms a low-resolution spectro-temporal pattern. The performance is comparable to the computationally more expensive AMS-based approach by Tchorz et al. and all algorithms tested by Ris and Dupont (still the time scales are different).

By definition the perceptual feature based SNR estimator is applicable for medium and long time scales (from 300ms up to several seconds) allowing noise reduction only for noise with slowly fluctuating components. Its main application could be automated selection of alternative hearing aid algorithms (compression, noise reduction) and classification of the acoustical environment. In addition, it is also useful in the context of robust ASR to control SNR dependent strategies on all processing levels (model compensation, choice of pre-processing or feature extraction schemes). The currently used integration over the complete segment length results in a delay of at least half the segment length, which is unfavorable for many applications. By replacing the integration of sigma-pi cells over a certain segment length with a leaky integrator, a continuous SNR estimation can be implemented without increasing the computational effort too much (a continuous estimation was already assumed when assessing the computational effort in Table 5.5).

The performance is better on relatively stationary types of noise and analysis shows highest estimation errors for modulation frequencies between 3 and 12Hz, which is identical to the range of modulation frequencies for speech, both broad band and in individual frequency bands. The perceptual feature based SNR estimator is negatively affected by noise sources which exhibit a similar temporal (and probably also spectral) envelope characteristic as speech. For some applications this is not a disadvan-

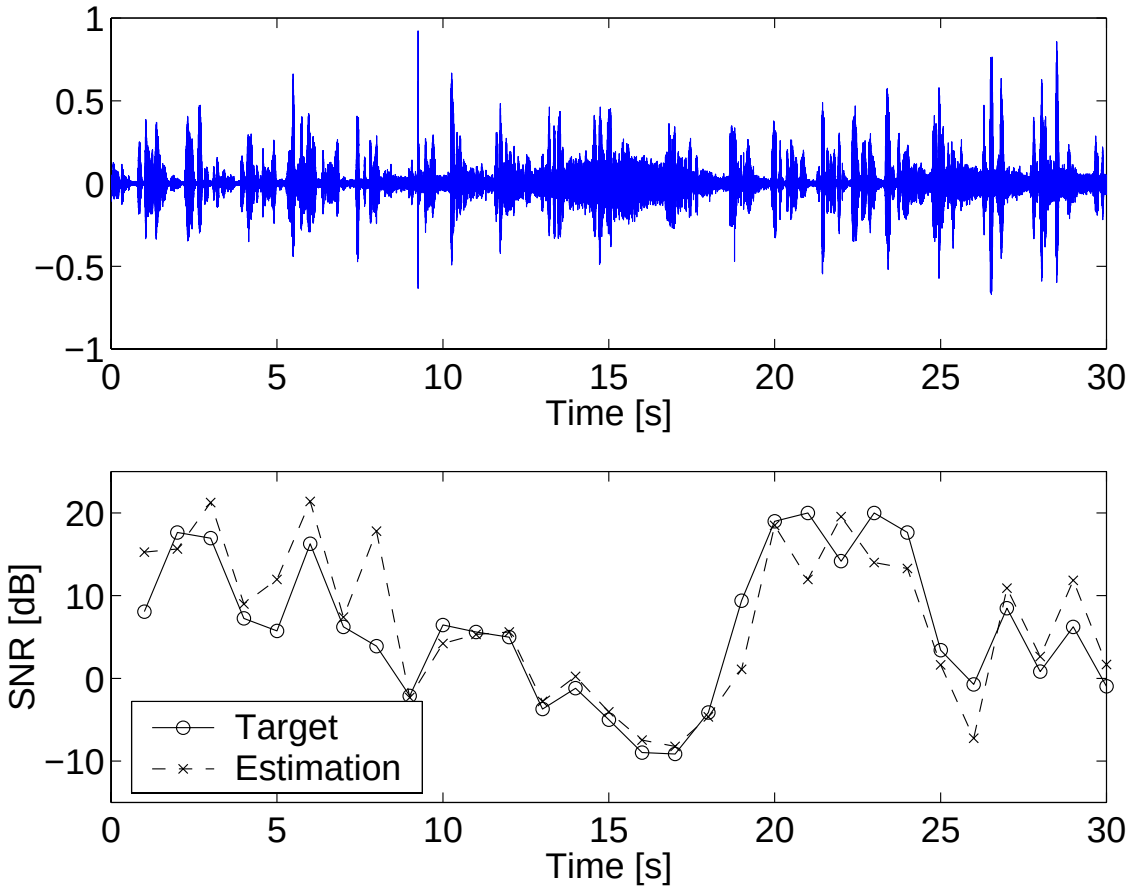


Figure 5.12: Mixed input signal (top) and estimated and true SNR values for channel five (center frequency 1296Hz, bottom). The noise example is from the test data set and features a jet engine plane passing by (labeled in the category of rather stationary noise).

tage. For example, the impact of noise reduction algorithms is reduced with overestimated SNR, which might help reducing processing artifacts in non-stationary types of noise, as many speech enhancement schemes cause distortions of speech in that case.

Although the training phase for neural networks requires some computational effort, once the network is trained the classification itself in the application phase is much easier to carry out. The back end sigma-pi/MLP combination requires relatively low computational effort and the somewhat expensive PEMO primary feature extraction might be replaced by other front ends if computing power is a constraint. The investigation of other front ends will be subject to future work. Another unwanted factor is the large difference in performance between training and test data. This problem might be overcome by training the algorithm on a larger database, which should also include a larger fraction of non-stationary noise signals.

To conclude, the perceptual feature based SNR estimator is capable of assessing the relative long-term levels of speech and background noise in the input signal with only little error for a large variety of realistic noise sources. This is achieved with reasonable computational load and therefore the proposed SNR estimator will be a beneficial addition for digital hearing aid and ASR systems.

Thanks to Birger Kollmeier for his substantial support and contribution to this work. Thanks also to Jürgen Tchorz for providing his compilation of noise and speech files for SNR estimation.

Further thanks to the anonymous reviewers and Stéphane Dupont for their help in improving the quality of this manuscript.

This work was supported by Deutsche Forschungsgemeinschaft (Project ROSE, Ko 942/15-1) and by Bundesministerium für Bildung und Forschung (Center of Excellence 'Hörtech', 01EZ0017).

chapter 6

Methods for Capturing Spectro-temporal Modulations in Automatic Speech Recognition [a]

Contents

6.1 Introduction 104
6.2 Secondary Features 107
6.3 Automatic Speech Recognition Experiments 114
6.4 Discussion 118

Abstract

Psychoacoustical and neurophysiological results indicate that spectro-temporal modulations play an important role in sound perception. Speech signals, in particular, exhibit distinct spectro-temporal patterns which are well matched by receptive fields of cortical neurons. In order to improve the performance of automatic speech recognition (ASR) systems a number of different approaches are presented, all of which target at capturing spectro-temporal modulations. By deriving secondary features from the output of a perception model the tuning of neurons towards different envelope fluctuations is modeled. The following types of secondary features are introduced: product of two or more windows (sigma-pi cells) of variable size in the spectro-temporal representation, fuzzy-logical (stochastic) combination of windows and a Gabor function to model the shape of receptive fields of cortical neurons. The different approaches are tested on

[a] A slightly different version of this chapter was published in *Acustica united with Acta Acustica* 88 (2002), pp. 416-422.

a simple isolated word recognition task and compared to a standard Hidden Markov Model recognition system. The results show that all types of secondary features are suitable for ASR. Gabor secondary features, in particular, yield a robust performance in additive noise, which is comparable and in some conditions superior to the Aurora 2 reference system.

Zusammenfassung

Ergebnisse aus Psychoakustik und Neurophysiologie weisen auf eine wichtige Rolle spektro-temporaler Modulationen für die auditorische Wahrnehmung hin. Insbesondere Sprachsignale zeigen deutliche spektro-temporale Muster, die gut zu den gefunden rezeptiven Feldern kortikaler Neurone passen. In diesem Kapitel werden eine Reihe von Ansätzen zur Verbesserung automatischer Spracherkennungsverfahren vorgestellt, die alle auf die Erfassung spektro-temporaler Modulationen zielen. Aus dem Ausgang der Perzeptionsmodels werden weitere, sekundäre Merkmale extrahiert, womit die Spezialisierung einzelner Neurone auf bestimmte Einhüllendenschwankungen modelliert wird. Folgende Arten sekundärer Merkmale werden verwendet: das Produkt zweier oder mehr Fenster (Sigma-Pi Zelle) variabler Größe, die Fuzzy-logische (stochastische) Kombination von Fenstern und schließlich die Gabor Filter Funktion als detaillierteres Modell der rezeptiven Felder von Neuronen. Die verschiedenen Ansätze werden für eine einfache Isoliertworterkennung untersucht und mit einem Standard Markovmodell Erkenner verglichen. Die Ergebnisse zeigen, daß alle drei Ansätze für die Spracherkennung geeignet sind. Insbesondere die Gabor-Merkmale erreichen eine robuste Erkennungsleistung in additiven Störgeräusch, vergleichbar denen des Referenzsystems und sind in einigen Fällen überlegen.

6.1 Introduction

Speech and many other natural sound sources exhibit distinct spectro-temporal amplitude modulations. While the temporal modulations are mainly due to the syllabic structure of speech, resulting in a bandpass characteristic with a peak around 4Hz (Kanedera et al., 1999), spectral modulations are due to the harmonic and formant structure of speech. The latter are not at all stationary over time. Coarticulation and intonation

result in variations of fundamental and formant frequencies even within a single phoneme (cf. Figure 6.1 as an example). The question is whether there is relevant information in amplitude variations oblique to the spectral and temporal axis and how it may be utilized to improve the performance of automatic classifiers.

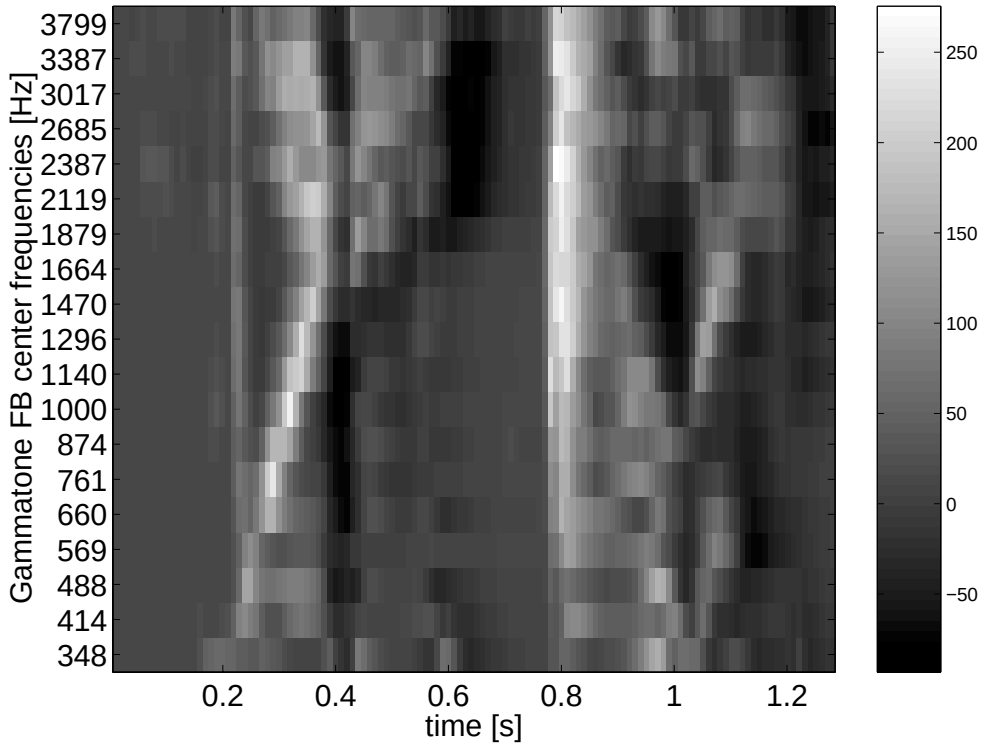


Figure 6.1: An example of a primary feature matrix for an utterance of the two words "Woody Allen" - in this case derived from the model of auditory perception as described in Section 6.3.2. Gray shading denotes output values in model units. A number of diagonal spectro-temporal structures may be identified.

In automatic speech recognition (ASR) the focus typically is on spectral modulation for a given time frame (cepstral analysis) *and/or* temporal fluctuations in individual frequency channels Hermansky and Morgan (1994); Hermansky and Sharma (1998). Although there are proposals to take two-dimensional variability into account (e.g. Weber et al., 2000)), auditory processing is not modeled explicitly.

Therefore, three different approaches are presented in this paper which target at capturing spectro-temporal modulations to increase the robustness of ASR systems:

Sigma-pi cells were originally proposed as a part of ASR systems in order to better capture certain features of speech like formants, formant

transitions, fricative onsets and (for larger units) phoneme sequences. A logical "AND" operation is performed by multiplicative combination of two spectro-temporal windows (Gramß and Strube, 1990). In Chapter 3, it was found that the combination of sigma-pi cells and auditory feature extraction yields very robust performance in additive noise. A generalization of this approach, towards a larger number of windows and variable window size, is motivated by recent psychoacoustical reverse correlation experiments. Using short segments of semi-periodic white Gaussian noise as stimuli, *early auditory features* of certain spectro-temporal shape were revealed (Kaernbach, 2000). These findings correspond well to physiological measurements of spectro-temporal receptive fields of neurons in the primary auditory cortex (deCharms et al., 1998) which often encompass different unconnected but highly localized parts of the spectrogram.

Fuzzy logic units: Due to its linear nature, the reverse correlation method does not reveal, if there has to be energy in regions A *and* B in order to stimulate a response or whether the receptive field is simply fragmented. To take account of this ambiguity the sigma-pi cell approach is extended to other fuzzy logical combination of windows, adding OR, NOR and NAND to the multiplicative AND operation.

Gabor functions are localized sinusoids and known to model the receptive fields of certain neurons in the visual system (De-Valois and De-Valois, 1990). In addition, experiments on human spectro-temporal modulation perception were modeled well by assuming a response field similar to two-dimensional Gabor functions (Chi et al., 1999). Therefore, in the third approach of this paper, two-dimensional Gabor receptive fields are examined for ASR. A complex two-dimensional Gabor function is calculated and reduced to real values by using only the real or imaginary component.

In the following the three types of secondary features are introduced and then applied to a simple isolated word recognition task for a first evaluation. Because of the large number of possible parameter combinations for all three variants of secondary features, the selection of a suitable sub-set is a major concern and the key to good classification performance. The classification and feature selection scheme described in Section 6.3.3 allows to automatically optimize a sub-set from all possible secondary features on

a given task and is therefore favored over standard ASR back ends in this approach.

6.2 Secondary Features

The secondary features $s_1(t)..s_M(t)$ are calculated from the primary feature values $p(t, f)$, which form a spectro-temporal representation of the input signal. t and f denote time and frequency channel index, respectively. The simplest examples of such two-dimensional representation (amplitude over frequency and time) are the spectrogram obtained by short-term Fourier analysis of consecutive time windows or, alternatively, a bank of band-pass filters. For speech and signal classification purposes, auditory-based approaches are likely to be more appropriate.

6.2.1 Sigma-pi Cells

Sigma-pi cells are known as second order elements from artificial neural network theory. This term describes certain network units in which the weighted outputs from two or more other units are multiplied before summation over all input values.

In the approach presented here, a number of windows $k \in [1, K]$ are defined centered around one element of the primary feature representation, which is located at frequency channel f_k and by t_k time steps shifted relative to the current feature vector. The windows have the extension Δt_k and Δf_k in time and frequency.

First, the average value w_k of each window is derived by

$$w_k = \frac{1}{\Delta t_k \Delta f_k} \sum_{t'} \sum_{f'} p(t_0 + t_k + t',\ f_k + f') \tag{6.1}$$

with $-\frac{\Delta t_k}{2} \leq t' \leq \frac{\Delta t_k}{2}$ and $-\frac{\Delta f_k}{2} \leq f' \leq \frac{\Delta f_k}{2}$.

The resulting value of any sigma-pi cell for time frame t_0 is then obtained from the window averages by:

$$s_m(t_k, f_k, \Delta t_k, \Delta f_k, t_0) = \prod_{k=1}^{K} w_k \tag{6.2}$$

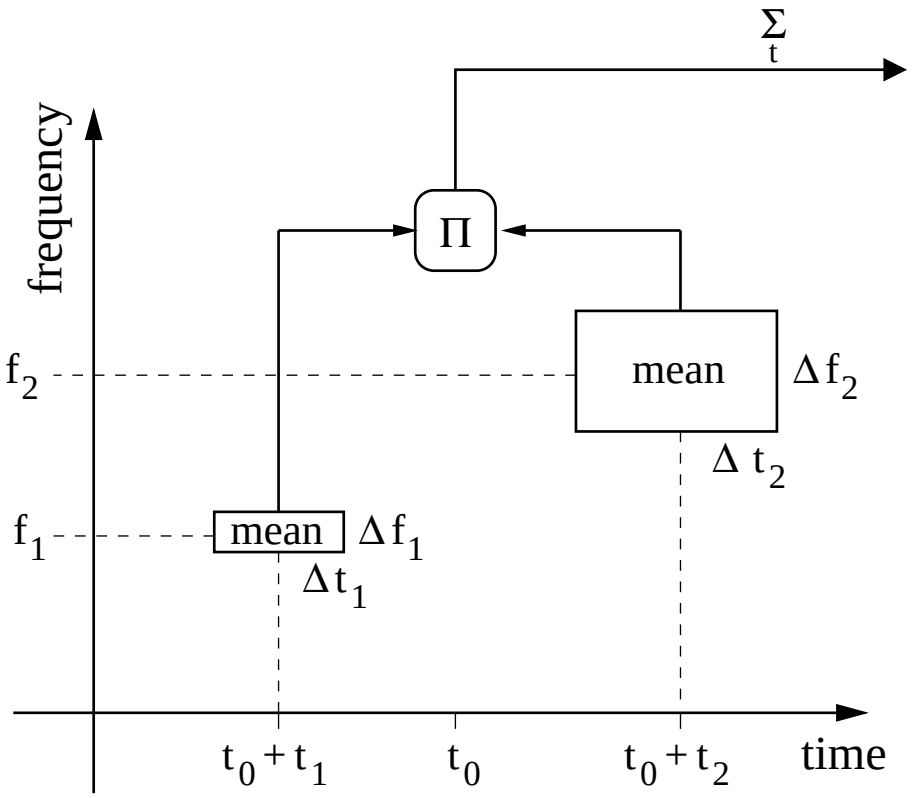


Figure 6.2: This sketch shows the denotation of parameters for a sigma-pi cell with two windows. See text for further description.

The secondary feature values $s_m(t_0)$ are often averaged over the whole utterance to obtain a single value per sigma-pi cell. Gramß and Strube (1990) proposed sigma-pi cells to be used as secondary features based on critical band spectrograms for isolated word recognition. Sigma-pi cells have later been used in combination with a perception model as front end for isolated word recognition and it was shown that this combination increases the robustness of ASR systems in additive noise (Chapter 3). With a non-linear back end the combination of perception model and sigma-pi cells is also suitable for sub-band signal-to-noise ratio (SNR) estimation (Chapter 5). In all those applications only two windows were used per sigma-pi cell and the smaller window was restricted to a single element of $p(t, f)$.

In the experiments presented below the window parameters for sigma-pi cells have the following constraints: $t_k \in [-20, 20]$ $(-200 - 200ms)$, $\Delta t_k \in [110]$ $(10 - 100ms)$, $\Delta f_k \in [1, 5]$ (ERB$^{\text{b}}$), and the number of windows $K \in 2, 3$. Furthermore, the windows have to be non-overlapping. Summation over time is performed to obtain a single secondary feature value per utterance.

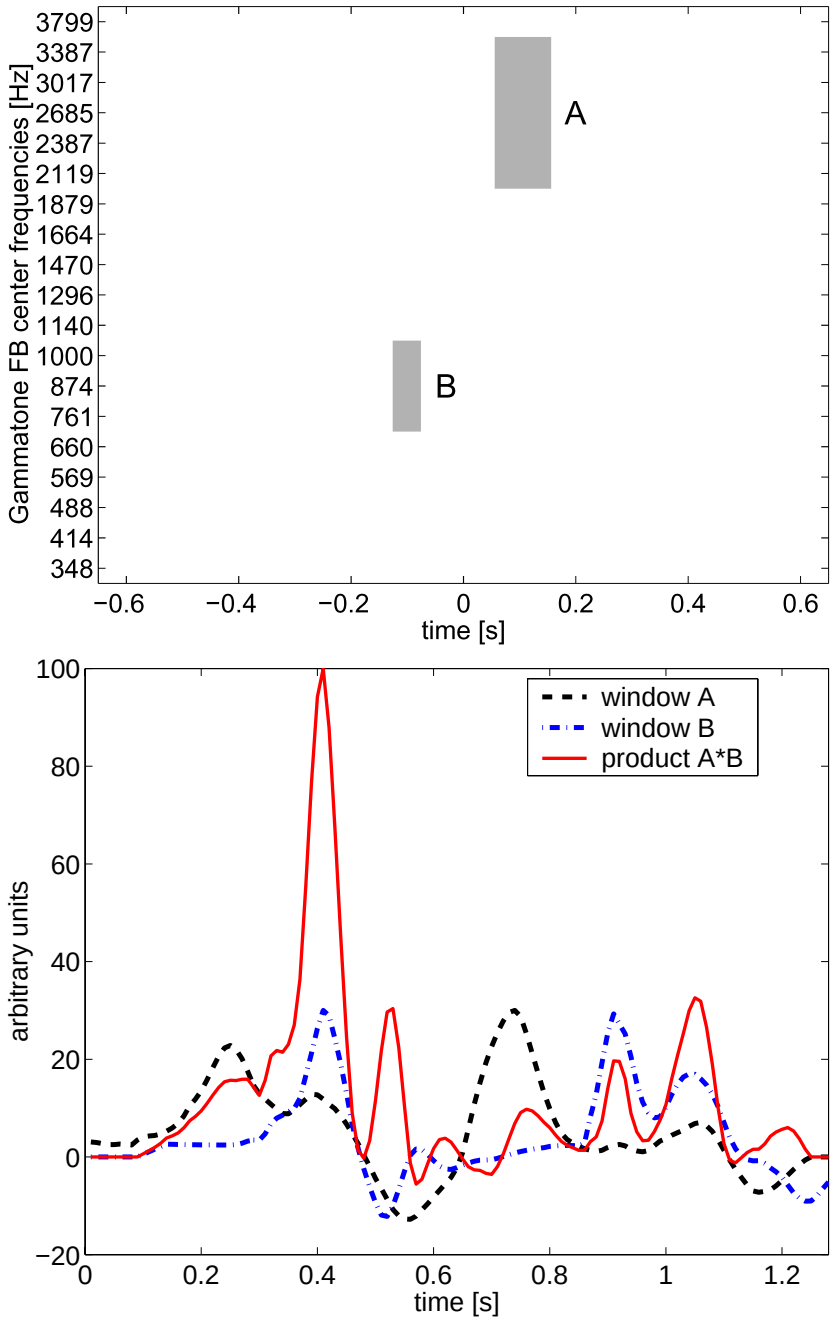


Figure 6.3: TOP: An example of a sigma-pi cell with two windows. Window A parameters are: $t = -10$ $(-100ms)$, $f = 7$ (ERB), $\Delta t = 5$ $(50ms)$ and $\Delta f = 3$ (ERB). Window B parameters are: $t = 10$ $(100ms)$, $f = 16$ (ERB), $\Delta t = 10$ $(100ms)$ and $\Delta f = 5$ (ERB). BOTTOM: Window averages and product of the two windows as a function of time, when the above sigma-pi cells is applied to the utterance depicted in Figure 6.1. The combination of the vowels /u/ and /i/ (or the lower and higher formants, respectively) in "Woody" was detected by the sigma-pi cells, by yielding large feature values around 0.4s.

Figure 6.3 gives an example on how a sigma-pi may serve as a feature detector. The sigma-pi cell is tuned to a sequence of phonetic elements in that case. The two windows, when coinciding with peaks in the spectro-temporal primary feature representation, basically detect spectro-temporal modulation of the frequency corresponding to the distance between the two windows. The temporal and spectral extension of the windows compensate to some degree for the variability inherent to spoken language. By calculating the product of the two windows, the secondary feature is of second order and the detection information remains even after integration over the whole time span of a word.

6.2.2 Fuzzy Logic Units

The sigma-pi cell approach is now extended by using true fuzzy logical (stochastic) combinations of windows instead of a simple multiplication, which corresponds to a logical AND. To obtain a value range between zero and one, the primary feature vectors are normalized by a logistic mapping function over the whole utterance:

$$p'(t,f) = \frac{1}{1+\exp\left[-\frac{p(t,f)-50}{25}\right]}. \tag{6.3}$$

or, alternatively, by a linear min-max normalization scheme:

$$p'(t,f) = \frac{p(t,f)-\min(p)}{\max(p)-\min(p)}. \tag{6.4}$$

The window averages w_k are calculated as in Eq. 6.1. The resulting value of a fuzzy logic unit for time t_0 is obtained recursively by:

$$s_{m,1}(t_0) = W_1(w_1) \tag{6.5}$$

and

$$s_{m,k}(t_0) = s_{m,k-1}\ O_{k-1}\ W_k(w_k). \tag{6.6}$$

The recursion terminates after K steps and the value $s_{m,K}$ is than adopted as secondary feature value $s_m = s_{m,K}$ for time t_0. The window operator

W_k is either identity ($f(A) = A$) or fuzzy complement (NOT operation), which is defined as $f(A) = 1 - A$. The possible fuzzy operators O_l are

intersection $f(A, B) = min(A, B)$

algebraic product $f(A, B) = A \cdot B$

union $f(A, B) = max(A, B)$

algebraic sum $f(A, B) = A + B - A \cdot B$.

The first two operators represent a fuzzy logical AND while the latter two correspond to fuzzy logical OR. With two or more windows a variety of combinations are possible. The NAND operation ('A AND NOT B'), for example, is assumed to be useful for edge detection in any spectro-temporal direction, while the AND operation ('A AND B', 'A AND NOT B AND C') serves as a detector for spectro-temporal modulations.

In the experiments described below, for fuzzy logic units the same parameter constraints applied as for sigma-pi cells.

6.2.3 Gabor Receptive Fields

The receptive field of cortical neurons is modeled as a two-dimensional complex Gabor function $g(t, f)$ defined as the product

$$g(\cdot) = n(\cdot) \cdot e(\cdot) \tag{6.7}$$

of the Gaussian envelope $n(t, f)$ with parameters $f_0, t_0, \sigma_f, \sigma_t$

$$n(\cdot) = \frac{1}{2\pi\sigma_x\sigma_t} \cdot \exp\left[\frac{-(f - f_0)^2}{2\sigma_f^2} + \frac{-(t - t_0)^2}{2\sigma_t^2}\right] \tag{6.8}$$

and the complex Euler function $e(t, f)$ with parameters $f_0, t_0, \omega_f, \omega_t$

$$e(\cdot) = \exp\left[i\omega_f(f - f_0) + i\omega_t(t - t_0)\right] \tag{6.9}$$

by using either the real or imaginary component only. The envelope width is defined by standard deviation values σ_f and σ_t. These are chosen as $\sigma = \frac{1}{\omega} \implies \sigma = \frac{T}{2\pi}$ for the imaginary component to ensure that only one

period of the oscillation gives a significant contribution to the function, and as $\sigma = \frac{\pi}{\omega} \implies \sigma = \frac{T}{2}$ for the real component. In the latter case the chosen combination of spread and periodicity leads to about 2.5 periods of the oscillation in the envelope and results in a negligible bias because

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\cdot)\ \mathrm{dt}\ \mathrm{df} \leq \exp\left[-\frac{\omega_t^2\sigma_t^2 + \omega_x^2\sigma_x^2}{2}\right] \tag{6.10}$$

and, with $\sigma_t = \frac{\pi}{\omega_t}$ and $\sigma_f = \frac{\pi}{\omega_f}$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\cdot)\ \mathrm{dt}\ \mathrm{df} \leq \exp\left[-\pi^2\right]. \tag{6.11}$$

This is important, because otherwise any stationary background signal would contribute to the secondary feature value.

In the experiments below the allowed temporal modulation frequencies $\frac{\omega_t}{2\pi}$ are limited to a range of one to 30Hz and the spectral modulations $\frac{\omega_f}{2\pi}$ to a range of 0.05 to 0.3 cycl/ERB, roughly corresponding to 0.25 - 1.5 cycl/oct. For a one ERB spectral resolution of the primary features, spectral modulations may only be calculated up to 0.5 cycles/ERB.

In order to extract a secondary feature value, the correlation between Gabor receptive field and the primary feature matrix is calculated. This matched filter operation is carried out in each frequency channel and the resulting values are summarized over all channels to obtain the activation $a(t_0, f_0, \omega_f, \omega_t, \sigma_f, \sigma_t)$ for each time step t_0. The cell response or secondary feature value for the whole utterance is then calculated as follows:

$$s_m(f_0, \omega_f, \omega_t, \sigma_f, \sigma_t) = \sum_{t_0=1}^{T} T\ [\ a(t_0)\ ] \tag{6.12}$$

with the non-linear transformation function T by either full-wave or half-wave rectification of $a(t_0)$.

In the experiments presented below, the primary feature vector sequence $p(t, f)$ is used either without or with min-max normalization (Eq. 6.4).

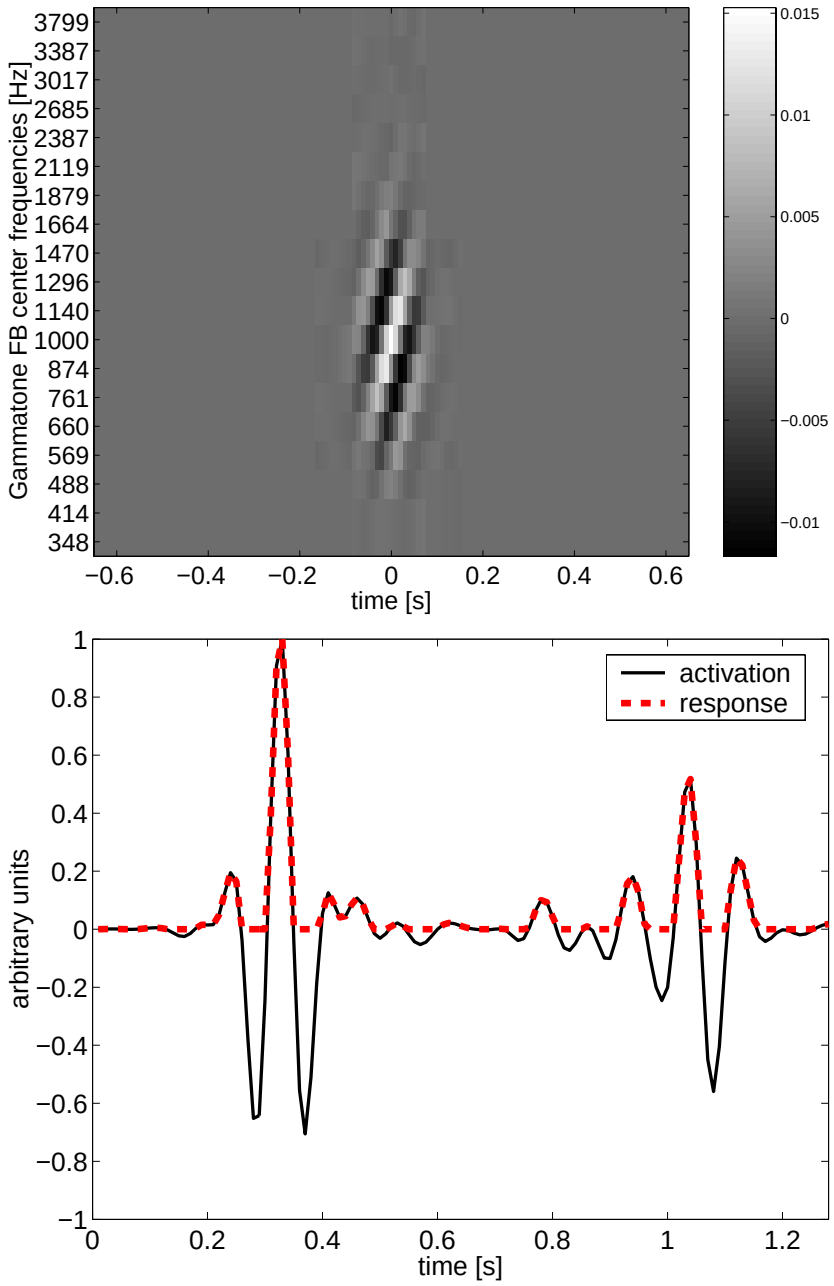


Figure 6.4: TOP: Example of the real component of a 2D Gabor function spectrally centered at 1000 Hz. Function values are given in shadings of gray. The Euler frequencies are $\frac{\omega_t}{2\pi} = -12Hz$ and $\frac{\omega_f}{2\pi} = 0.2$cycles/channel. The function is calculated on a grid with 100 Hz temporal and 1/ERB spectral sampling, according to the primary feature extraction method used in this study. BOTTOM: Filter output ("activation") and halfway rectified feature values ("response") over time when the above Gabor filter is applied to the utterance depicted in Figure 6.1. The rising formant between 0.3 and 0.4s fits the Gabor filter shape well and yields highest feature values. A similar diagonal feature is detected around 1.1s, resulting in a second, somewhat smaller peak.

While the imaginary component might be able to serve as edge detector in the spectro-temporal domain, the real component is designed to capture spectro-temporal modulations in any possible direction - including simple temporal or spectral modulations. The wide range of possible Gabor features is therefore versatile enough to contain purely spectral features (as cepstra) or temporal processing (as in the RASTA or TRAPS approaches). The above mentioned front ends are extended as most of the possible Gabor filters perform integrated spectral *and* temporal processing. Figure 6.4 shows one example of such a diagonal Gabor feature function and how it can be used to detect formant transitions.

6.3 Automatic Speech Recognition Experiments

6.3.1 Material

The speech material for training and testing is taken from the ZIFKOM database[c]. Each German digit was recorded once from 200 different speakers. The speech material is equally divided into two parts for training and testing, each consisting of 1000 utterances by 50 male and 50 female speakers. Training is performed on clean digits only. Testing is performed on clean and on noisy digits. For distortion, three types of noise are added to the utterances with SNR between 25 and -5dB: a) un-modulated speech shaped noise (CCITT G.227), with a spectrum similar to the long-term spectrum of speech, b) real babble noise recorded in a cafeteria situation and c) speech-like shaped and modulated noise (ICRA noise signal 7, ICRA, 1997)[d]. Before mixing, speech and noise signals are bandpass filtered to 300-4000Hz, roughly corresponding to the telephone band.

6.3.2 Primary Feature Extraction

The output of the model of auditory perception (PEMO) is used as primary feature matrix. PEMO has been originally developed by Dau et al. (1996a) for quantitatively simulating psychoacoustical experiments, such as temporal and spectral masking, and has been successfully applied as

[c]Deutsche Telekom AG

[d]two foreground speakers and four background speakers

a robust front end in isolated word recognition experiments (Tchorz and Kollmeier, 1999a, Chapter 2). Its major components are the peripheral gammatone filter bank (Hohmann, 2002) and the non-linear adaptation loops (Püschel, 1988), which perform a log-like compression for stationary signals and emphasize onsets and offsets of the envelope. This causes a sparse coding of the input in the spectro-temporal domain. It should be stressed that any other time-frequency amplitude representation could also be used with this approach, preferably an auditory model or auditory-like processing (Chapter 5).

In this study, the model was slightly modified by adding a pre-emphasis[e], which is motivated by earlier ASR experiments (Chapter 3). Overall, 19 frequency channels are used with bandwidth and spacing of one ERB and center frequencies ranging from 384 to 3799Hz. The primary feature vectors are then derived by downsampling the model output to a sampling frequency of $f_s = 100$Hz in each channel.

6.3.3 Recognizer

For classification and optimization of the type of secondary features the *Feature-finding Neural Network* (FFNN) Gramß and Strube (1990) is used. It consists of a linear single-layer perceptron in conjunction with secondary feature extraction and an optimization rule for the feature set. For a sufficiently high-dimensional feature space (i.e. a large number of secondary features), a linear net should classify equally well as non-linear classifiers and fast training is guaranteed by matrix inversion (pseudo-inverse method). Given P examples, each represented by a secondary feature vector with M elements, the feature vectors form a $M \times P$ feature matrix $\mathbf{X}$. Given the target matrix $\mathbf{Y}$ ($N \times P$ with N as the number of classes or target values per example), the optimal (in RMS sense) weight matrix $\mathbf{W}$ ($N \times M$) is found analytically by calculating the pseudo-inverse

$$\mathbf{X}^+ = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{6.13}$$

of the secondary feature matrix $\mathbf{X}$. The weight matrix is obtained as

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^+ \tag{6.14}$$

[e]differentiation with factor of 0.97: $y_n = x_n - 0.97 \cdot x_{n-1}$

and minimizes the classification error

$$E = |\mathbf{Y} - \mathbf{W}\mathbf{X}|^2. \tag{6.15}$$

Gramß Gramß (1991) proposed a number of training algorithms for the FFNN system, one of which, the *substitution rule*, is used in this study:

i. Choose M secondary features arbitrarily.

ii. Find the optimal weight matrix $\mathbf{W}$ using all M features and the M weight matrices that are obtained by using only $M-1$ features, thereby leaving out every feature once.

iii. Measure the relevance R of each feature i by

$$R_i = E(\text{without feature } i) - E(\text{with all features}) \tag{6.16}$$

iv. Discard the least relevant feature $j = \text{argmin}(R_i)$ from the sub-set and randomly select a new candidate.

v. Repeat from point 2. until the maximum number of iterations is reached.

vi. Recall the set of secondary features that performed best on the training/validation set and return it as result of the substitution process (modification from original substitution rule).

Although the classification is performed by a linear neural network, the whole classification process is highly non-linear due to the second order characteristics of the secondary features. The thereby obtained set of secondary features might also be used as input to other, more sophisticated classification systems. The segmentation problem is not relevant for an isolated word recognition task and therefore the summation of secondary feature values over the whole utterance is a sufficiently good option to derive a single value per secondary feature and utterance. In the more general continuous case, e.g., a leaky integrator could be used to extract time-depending secondary feature values.

In the experiments below, a set of 60 secondary features is optimized over 2000 iterations. Due to the non-deterministic nature of the substitution rule (random start set and randomly chosen substituting secondary feature), training is carried out eight times per configuration.

6.3.4 Results

The results are summarized in Table 6.1. All three types of secondary feature are suitable for ASR. Gabor features perform best in CCITT noise and on clean test material and comparable to sigma-pi cells for babble and ICRA 7 noise. Fuzzy logic secondary features lead to an unacceptable high error for clean test data and also to the highest word error rate (WER) values in most other cases. The robustness of fuzzy logic features can be increased by using min-max normalization instead of logistic function (Table 6.2), but the error rate for clean data remains too high also in that case.

Table 6.1: Word error rates (WER) in percent for different SNR (in dB) and noise conditions. 'train' indicates the training material, while 'clean' refers to the unmixed test data. Mean and standard deviation (in brackets) over 8 training runs per condition are given for sigma-pi cell, fuzzy logic (logistic normalization) and Gabor secondary features.

condition	SNR [dB]	Sigma-pi		Fuzzy (logistic)		Gabor	
train		0.5	(0.2)	1.0	(0.2)	0.4	(0.2)
clean		2.0	(0.3)	3.3	(0.6)	1.1	(0.2)
CCITT	25	4.9	(1.2)	9.0	(2.7)	5.1	(1.2)
	20	11.7	(2.0)	22.2	(7.4)	11.1	(3.1)
	15	35.3	(4.0)	47.9	(10.9)	27.5	(8.6)
	10	67.1	(4.8)	72.3	(6.1)	52.7	(9.9)
	5	82.8	(5.2)	83.5	(3.4)	72.0	(5.5)
	0	88.5	(1.7)	88.2	(1.2)	82.3	(3.8)
	-5	89.8	(0.3)	89.6	(0.5)	87.2	(2.3)
BABBLE	25	3.6	(0.7)	8.2	(0.8)	4.5	(1.0)
	20	6.3	(1.6)	16.3	(2.3)	8.6	(2.7)
	15	16.9	(3.5)	33.5	(7.8)	22.2	(7.4)
	10	43.0	(4.7)	54.5	(11.2)	45.8	(10.5)
	5	68.1	(4.6)	72.0	(7.8)	68.0	(9.0)
	0	82.4	(3.9)	82.1	(3.4)	81.3	(4.8)
	-5	87.4	(2.4)	87.5	(2.3)	87.5	(2.1)
ICRA 7	25	3.6	(0.7)	7.4	(1.1)	4.0	(1.1)
	20	6.6	(1.3)	15.1	(3.4)	9.0	(4.0)
	15	17.2	(4.1)	30.7	(7.1)	23.5	(11.7)
	10	44.5	(6.3)	51.6	(9.8)	46.1	(18.3)
	5	70.9	(3.2)	70.5	(8.0)	66.4	(17.5)
	0	83.0	(2.5)	80.9	(5.3)	78.3	(12.8)
	-5	87.9	(1.9)	86.4	(2.7)	84.3	(7.4)

Gabor receptive fields yield lower WER values than sigma-pi cells in most cases. This is remarkable, because the Gabor secondary features are of 1st

order, while the other two variants are 2nd order features. The variance of performance over different training runs is relatively high, especially for Gabor receptive fields in the case of additive speech-like modulated noise (ICRA 7). As the optimization is carried out on clean training data, only in some cases the secondary features seem to be affected by the modulation in the noise signal (which is kept frozen for all examples). In Table 6.2 WER for the most robust single set of Gabor features out of eight sets are shown. The large variance of WER in noise between the eight sets of optimized Gabor secondary features indicate that some sets of Gabor receptive fields contain features which are less suitable in noisy conditions. Multi-condition training is likely to increase the robustness by selecting only noise-robust type of features into the optimal set.

As a reference, the Aurora 2 baseline system Hirsch and Pearce (2000) has been applied to the same classification task. It is composed out of the WI007 (mel-cepstrum) front end and a reference HTK recognizer. The results obtained by this Hidden Markov Model classifier are presented in Table 6.2 and compared to improved Gabor secondary features.

Both, the best Gabor set of secondary features and Gabor secondary feature set with min-max normalization of primary feature values, show a comparable robustness to the aurora baseline system on the given classification task. There is a trend for the aurora system to yield lower WER for clean test data and high SNR values of over 10dB while the Gabor secondary features seem to be superior in more unfavorable conditions of low SNR values. It should be stressed that the classifier used here for the secondary features is as simple as possible with a summation over the whole utterance followed by a linear neural network. Therefore, an increase in performance can be expected when combining time-dependent secondary features, e.g., Gabor receptive fields, with a more sophisticated classifier.

6.4 Discussion

The proposed extensions to the secondary feature approach are all suitable for for robust isolated word recognition. Especially the Gabor receptive field method seems to be worthwhile to be investigated further. Gabor secondary features combined with a simple linear classifier show a comparable performance to the state-of-the-art Aurora 2 HMM system. They can be assumed to have a large potential. Earlier studies indicate, for

Table 6.2: Word error rates (WER) in percent for different SNR (in dB) and noise conditions. 'train' indicates the training material, while 'clean' refers to the unmixed test data. Mean and standard deviation (in brackets) over 8 training runs per condition are given for fuzzy logic units and Gabor receptive fields - both with min-max normalization of primary feature vectors. The most robust single set of Gabor features without normalization ('Gab. best') is compared to the Aurora 2 baseline system ('Aurora'), which is given as a reference.

condition	SNR [dB]	Fuzzy (min-max)		Gabor (min-max)		Gabor best	Aurora
train		0.5	(0.1)	0.3	(0.1)	0.5	0.3
clean		3.7	(0.6)	1.7	(0.3)	1.1	0.3
CCITT	25	4.6	(1.0)	3.8	(0.8)	4.6	1.7
	20	6.8	(1.4)	5.8	(1.8)	7.6	3.9
	15	12.9	(2.9)	12.0	(4.1)	16.7	9.7
	10	29.2	(6.2)	26.8	(8.5)	37.9	24.1
	5	51.8	(9.2)	50.1	(11.5)	66.4	73.8
	0	69.8	(8.2)	73.0	(10.0)	80.5	90.9
	-5	81.3	(5.7)	85.4	(5.3)	85.0	90.6
BABBLE	25	4.4	(0.6)	3.4	(0.5)	3.4	1.2
	20	6.1	(1.0)	5.2	(1.1)	4.8	2.3
	15	10.7	(1.1)	10.3	(2.5)	9.0	4.1
	10	21.9	(2.5)	22.4	(5.2)	22.7	14.1
	5	42.5	(5.1)	43.4	(5.5)	46.6	42.0
	0	65.9	(6.6)	64.9	(6.0)	70.3	72.6
	-5	82.0	(4.9)	80.0	(3.5)	83.0	83.5
ICRA 7	25	4.6	(0.9)	2.8	(0.5)	2.8	1.1
	20	7.6	(1.4)	4.7	(0.9)	3.8	1.6
	15	14.6	(2.2)	9.4	(2.8)	7.4	4.0
	10	28.3	(3.9)	20.3	(6.0)	15.5	14.8
	5	48.0	(4.6)	38.9	(7.6)	30.2	31.3
	0	67.8	(2.6)	59.8	(5.3)	50.7	54.8
	-5	81.3	(2.3)	75.6	(4.9)	69.1	83.7

example, an increase in robustness equivalent to a five to eight dB effective gain in SNR by using noise reduction pre-processing schemes with PEMO primary features (Chapter 2). Classification performance should increase further by replacing the simple linear network classifier with a state-of-the-art HMM back end and/or adding spectro-temporal features as another feature stream in a multi-stream system.

The author would like to thank Volker Hohmann and Birger Kollmeier for their substantial support and contribution to this work. Thanks also to

Christian Kaernbach for stimulating conversation and his idea to use fuzzy logic, to Heiko Gölzer for fruitful discussion about optimization rules.

This work was supported by Deutsche Forschungsgemeinschaft (Project ROSE, Ko 942/15-1).

CHAPTER 7

Spectro-temporal Gabor Features as a Front End for Automatic Speech Recognition [a]

Contents

7.1 Introduction 122
7.2 Gabor Filter Functions 124
7.3 Feature Selection 126
7.4 ASR Experiments 132
7.5 Summary 133

Abstract

A novel type of feature extraction is introduced to be used as a front end for automatic speech recognition (ASR). Two-dimensional Gabor filters are applied to a spectro-temporal representation of the input signal formed by columns of primary feature vectors. The filter shape is motivated by recent findings in neurophysiology and psychoacoustics which revealed sensitivity towards complex spectro-temporal modulation patterns. Supervised data-driven parameter selection yields qualitatively different feature sets depending on the corpus and the target labels. The overall distribution of temporal and spectral modulation frequencies in the sets reflects properties of speech. ASR experiments on the Aurora dataset show the benefit of the proposed Gabor features, especially in combination with other feature streams.

[a] A slightly different version of this chapter was published as a proceedings paper accompanying an invited talk at the *Forum Acusticum* in Seville 2002.

Zusammenfassung

In diesem Kapitel wird ein neuartiges Verfahren zur Merkmalsextraktion in der automatischen Spracherkennung eingeführt. Dabei werden zweidimensionale Gabor Filter auf eine spektro-temporale Repräsentation des Eingangssignals angewendet, welche aus Spalten von Merkmalsvektoren besteht. Die Filterform ist durch Ergebnisse neurophysiologischer und psychoakustischer Forschung motiviert, die eine spektro-temporale Verarbeitung nahelegen. Überwachte, datenbasierte Parameteroptimierung resultiert in qualitativ anderen Merkmalen für verschiedene Korpora und Spracheinheiten. Die Gesamtverteilung zeitlicher und spektraler Modulationsparameter spiegelt die Eigenschaften von Sprache wider. Spracherkennungsexperimente im Rahmen des Aurora Datensatzes zeigen den Vorteil der Gabor-Merkmale, insbesondere in Kombination mit anderen Merkmalsvektortypen.

7.1 Introduction

ASR technology has seen many advances in recent years, still the issue of robustness in adverse conditions remains largely unsolved. Additive noise as well as convolutive noise in the form of reverberation and channel distortions occur in most natural situations, limiting the feasibility of ASR systems in real world applications. Standard front ends, such as mel-cepstra or perceptual linear prediction, only represent the spectrum within short analysis frames and thereby neglect very important dynamic patterns in the speech signal. This deficiency has been partly overcome by adding temporal derivatives in the form of delta and delta-delta features to the set. In addition, channel effects can be reduced by carrying out further temporal bandpass filtering such as cepstral mean subtraction or RASTA processing (Hermansky and Morgan, 1994). A completely new school of thought has been initiated by a review of Fletcher's work (Allen, 1994), who found log sub-band classification error probability to be additive for nonsense syllable recognition tasks observed on human subjects. This suggests independent processing in a number of articulatory bands without recombination until a very late stage. The most extreme example of the new type of purely temporal features are the TRAPS (Hermansky and Sharma, 1998) which apply multi-layer perceptrons (MLP) to classify current phonemes in each single critical band based on a temporal context

of up to 1s. Another approach is multi-band processing (Bourlard et al., 1996a), for which features are calculated in broader sub-bands to reduce the effect of band-limited noise on the overall performance. All these feature extraction methods apply either spectral or temporal processing at a time. Nevertheless, speech and many other natural sound sources exhibit distinct spectro-temporal amplitude modulations (see Figure 7.2 a) as an example). While the temporal modulations are mainly due to the syllabic structure of speech, resulting in a bandpass characteristic with a peak around 4Hz, spectral modulations describe the harmonic and formant structure of speech. The latter are not at all stationary over time. Coarticulation and prosody result in variations of fundamental and formant frequencies even within a single phoneme. This raises the question whether there is relevant information in amplitude variations oblique to the spectral and temporal axes and how it may be utilized to improve the performance of automatic classifiers. In addition, recent experiments about speech intelligibility showed synergetic effects of distant spectral channels (Greenberg et al., 1998) that exceed the log error additivity mentioned earlier and therefore suggest spectro-temporal integration of information. This is supported by a number of physiological experiments on different mammal species which have revealed the spectro-temporal receptive fields (STRF) of neurons in the primary auditory cortex. Individual neurons are sensitive to specific spectro-temporal patterns in the incoming sound signal. The results were obtained using reverse correlation techniques with complex spectro-temporal stimuli such as checkerboard noise (deCharms et al., 1998) or moving ripples (Schreiner and Calhoun, 1994; Kowalski et al., 1996). The STRF often clearly exceed one critical band in frequency, have multiple peaks and also show tuning to temporal modulation (Schreiner et al., 2000). In many cases the neurons are sensitive to the direction of spectro-temporal patterns (e.g. upward or downward moving ripples), which indicates a combined spectro-temporal processing rather than consecutive stages of spectral and temporal filtering (Depireux et al., 2001). These findings fit well to psychoacoustical experiments on early auditory features (Kaernbach, 2000), yielding patterns that are distributed in time and frequency and in some cases comprised of several unconnected parts. These STRF can be approximated, although somewhat simplified, by two-dimensional Gabor functions, which are localized sinusoids known from receptive fields of neurons in the visual cortex (De-Valois and De-Valois, 1990).

In this paper, new two-dimensional features are investigated, which can be obtained by filtering a spectro-temporal representation of the input signal with Gabor-shaped localized spectro-temporal modulation filters. These new features in some sense incorporate but surely extend the features mentioned above. A recent study showed an increase in robustness when real-valued Gabor filters are used in combination with a simple linear classifier on isolated word recognition tasks (Chapter 6). Now, the Gabor features are modified to a complex filter and based on mel-spectra, which is the standard first processing stage for most types of features mentioned above. It is investigated whether the use of Gabor features may increase the performance of more sophisticated state-of-the-art ASR systems. The problem of finding a suitable set of Gabor features for a given task is addressed and optimal feature sets for a number of different criteria are analyzed.

7.2 Gabor Filter Functions

The Gabor approach pursued in this paper has the advantage of a neurobiological motivated prototype with only few parameters, which allows for efficient automated feature selection. The parameter space is wide enough to cover a large variety of cases: purely spectral features are identical to sub-band cepstra - modulo the windowing function - and purely temporal features closely resemble the TRAPS pattern or the RASTA impulse response and its derivatives (Hermansky, 1998). Gabor features are derived from a two-dimensional input pattern, typically a series of feature vectors. A number of processing schemes may be considered for these primary features that extract a spectro-temporal representation from the input wave form. The range is from a spectrogram to sophisticated auditory models. In this study the focus is on the log mel-spectrogram for its widespread use in ASR, and because it can be regarded as a very simple auditory model, with instantanious logarithmic compression and mel-frequency axis. Here, the log mel-spectrum was calculated as in the ETSI aurora standard and reference (ETSI ES 201 v1.1.2, 2000). The processing consists of DC removal, Hanning windowing with 10ms offset and 25ms length, pre-emphasis, FFT and summation of the magnitude values into 23 mel-frequency channels with center frequencies from 124 to 3657Hz. The amplitude values are then compressed by the natural logarithm. The receptive field of cortical neurons is modeled by two-dimensional complex Gabor functions $g(t, f)$

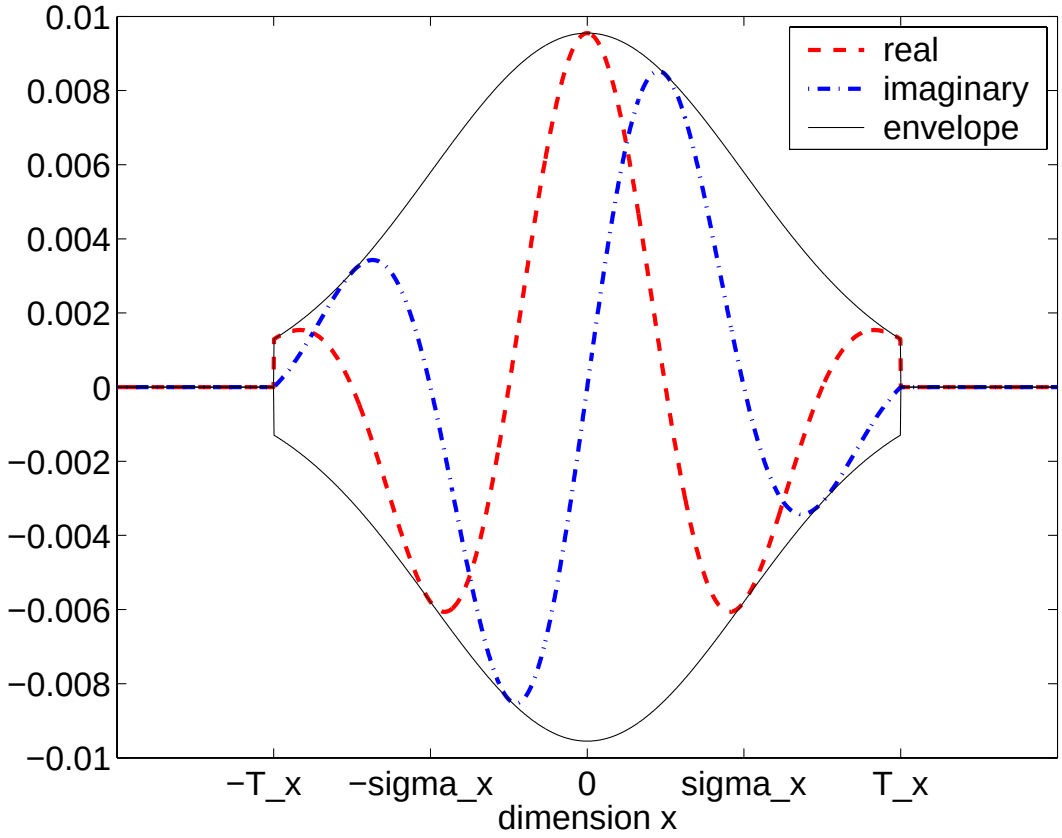


Figure 7.1: Example of a one-dimensional complex Gabor function or a cross section of a two-dimensional one. Real and imaginary components are plotted, corresponding to $\pi/2$ and zero phase, respectively. Note that one period $T_x = 2\pi/\omega_x$ of the oscillation fits into the interval $[-\sigma_x, \sigma_x]$ and the support in this case is reduced from infinity to twice that range or $2T_x$. An example of a 2D-Gabor function can be found in Figure 7.2 b.

defined as the product of a Gaussian envelope $n(t, f)$ and the complex Euler function $e(t, f)$. The envelope width is defined by standard deviation values σ_f and σ_t, while the periodicity is defined by the radian frequencies ω_f and ω_t with f and t denoting the frequency and time axis, respectively. Further parameters are the centers of mass of the envelope in time and frequency t_0 and f_0. In this notation the Gabor function $g(t, f)$ is defined as

$$g(t,f) = \frac{1}{2\pi\sigma_x\sigma_t} \cdot \exp\left[\frac{-(f-f_0)^2}{2\sigma_f^2} + \frac{-(t-t_0)^2}{2\sigma_t^2}\right] \cdot \exp\left[i\omega_f(f-f_0) + i\omega_t(t-t_0)\right]. \tag{7.1}$$

It is reasonable to set the envelope width depending on the modulation frequencies in order to keep the same number of periods in the filter func-

tion for all frequencies. Basically, this makes the Gabor feature a wavelet prototype with a scale factor for each of the two dimensions. The spread of the Gaussian envelope in dimension x was set to $\sigma_x = \pi/\omega_x = T_x/2$ to have a full period T_x in the range between $-\sigma_x$ and σ_x as depicted in Figure 7.1. The infinite support of the Gaussian envelope is cut off at σ_x to $2\sigma_x$ from the center. For time dependent features, t_0 is set to the current frame, so three main free parameters remain: f_0, ω_f and ω_t. The range of parameters is limited mainly by the resolution of the primary input matrix (100Hz and 23 channels covering 7 octaves). The temporal modulation frequencies were limited to a range of two to 50Hz, and the spectral modulation frequencies to a range of 0.04 to 0.5 cycles per channel or approximately 0.14 to 1.64 cycles per octave. If ω_f or ω_t is set to zero to obtain purely temporal or spectral filters, respectively, σ_t or σ_f again becomes a free parameter.

From the complex results of the filter operation, real valued features may be obtained by using the real or imaginary part only. This method was used earlier (Chapter 6) and offers the advantage of being sensitive to the phase of the filter output and thereby to the exact temporal and spectral location of events (cf. Figure 7.2 c and d). Alternatively, the magnitude of the complex filter output may be used. This gives a smoother filter response (cf. Figure 7.2 e and f) and allows for a phase independent feature extraction which might be advantageous in some cases. Both types of filters have been used in the experiments below. The filtering is performed by calculating the correlation function at all time delays of each input frequency channel with the corresponding part of the Gabor function and a subsequent summation over frequency. This yields one output value per frame per Gabor filter and is equivalent to a two-dimensional correlation of the input representation with the complete filter function and a subsequent selection of the desired frequency channel f_0 (see Figure 7.2).

7.3 Feature Selection

Due to the large number of possible parameter combinations, it is necessary to select a suitable set of features. This was carried out by a modified version of the Feature-finding Neural Network (FFNN). It consists of a linear single-layer perceptron in conjunction with secondary feature extraction and an optimization rule for the feature set (Gramß and Strube, 1990).

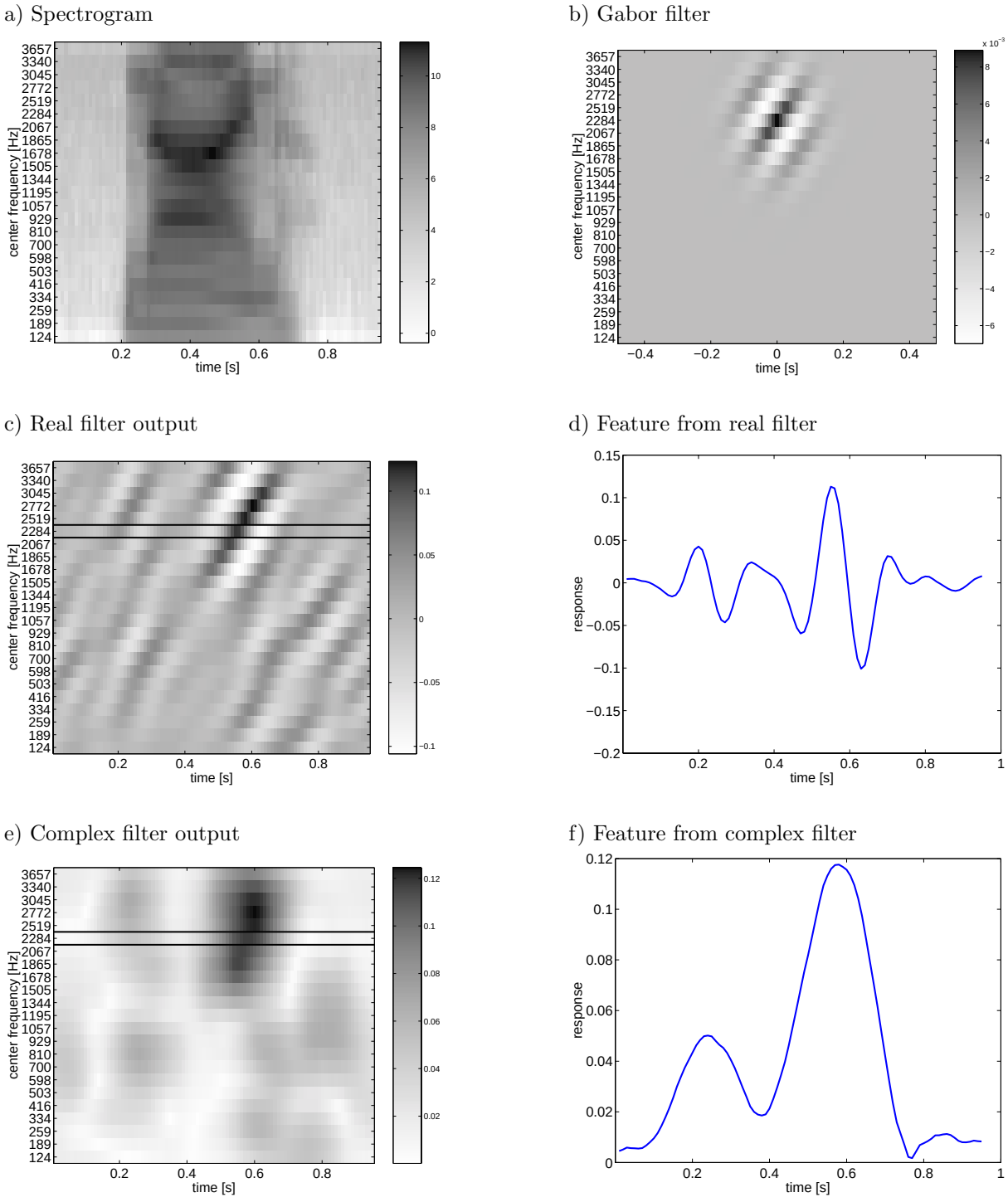


Figure 7.2: a) mel-scale log magnitude spectrogram of a "Nine" from the TIDigits corpus. b) an example of a 2D-Gabor complex filter function (real values plotted here) with parameters $\omega_t/2\pi$ = -7Hz and $\omega_f/2\pi$ = 0.2 cycl./channel. The resulting filtered spectrograms for c) real and e) complex valued filters. d) and f): The resulting feature values for f_0 = 2284Hz.

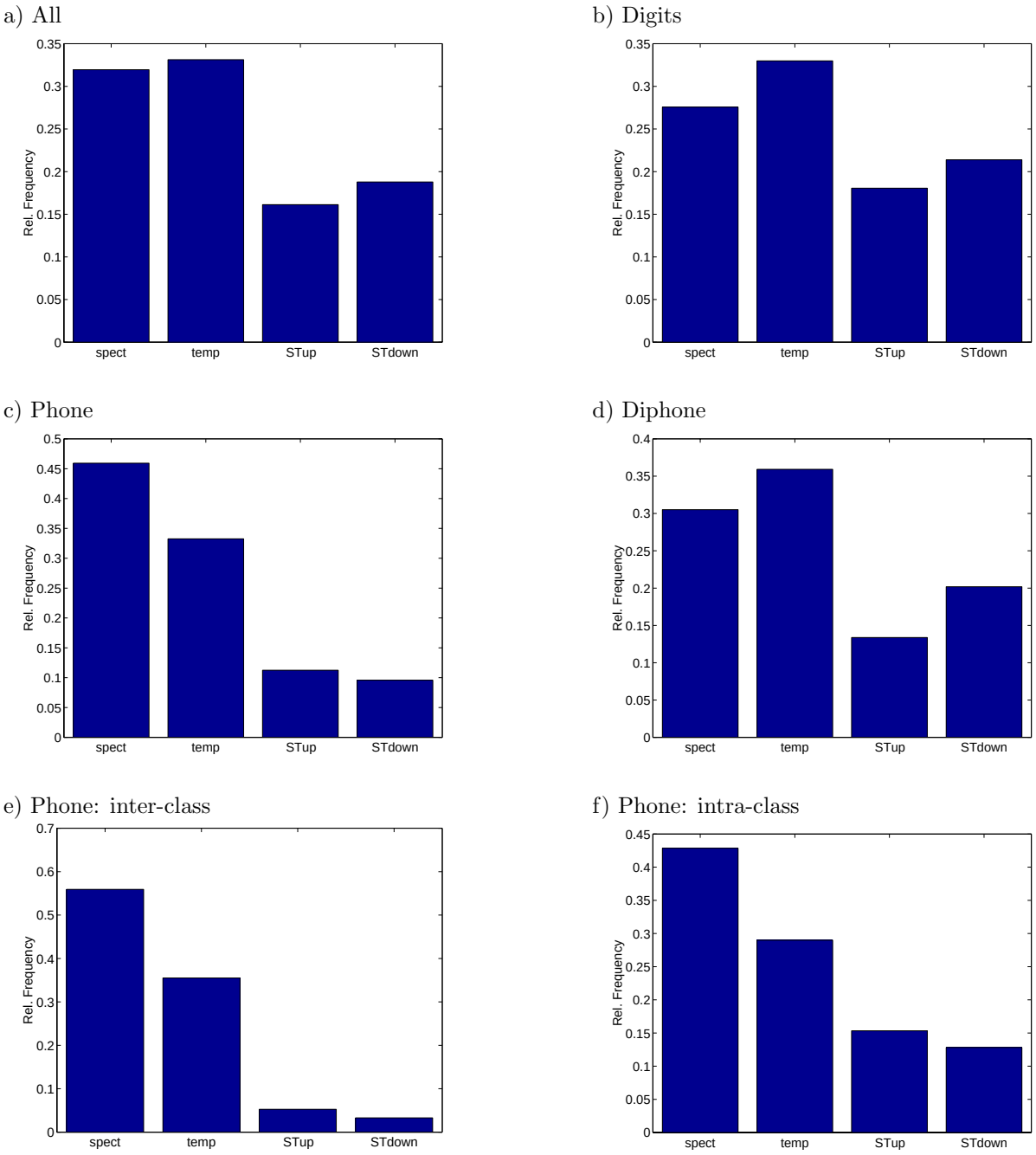
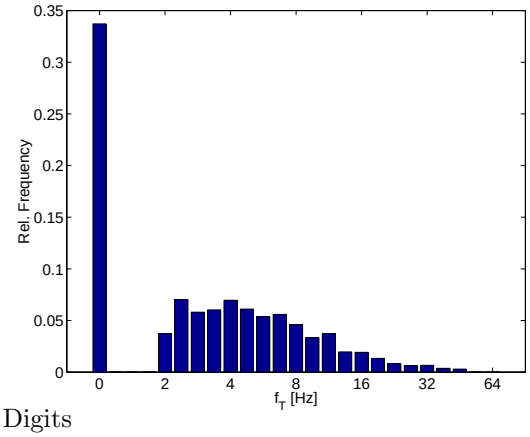
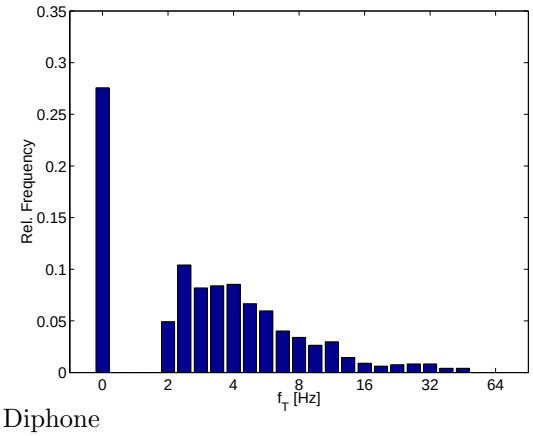


Figure 7.3: Distribution of Gabor types a) in all selected sets (103 sets with 2702 features) and b) for digits (43/1440), c) phone (38/836) and d) diphone (22/426) targets only. Overall percentages of spectral, temporal and spectro-temporal (ST) features are given. 'down' denotes negative temporal modulation frequency. Distribution of Gabor types for phone targets with grouping into e) broad phonetic (manner) classes (8/152) and f) for single phonetic classes (18/476).

a) All



b) Digits



c) Diphone

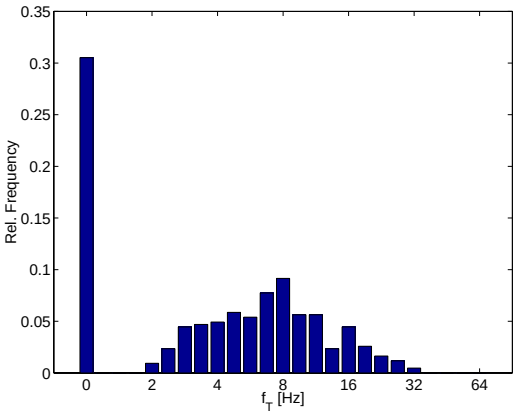


Figure 7.4: Distribution of temporal modulation frequency $\omega_t/2\pi$ over all Gabor types a) in all selected sets, b) for digits and c) for diphone targets. Purely spectral features accumulate in the 0Hz bin, although they also have a limited temporal extend. d) Distribution of spectral modulation frequency for all targets. Purely temporal features accumulate in the 0 bin, although they also have a limited spectral extend.

The linear classifier guarantees fast training, which is necessary because in this wrapper method for feature selection the importance of each feature is evaluated by the increase of RMS classification error after its removal from the set. This 'substitution rule' method (Gramß, 1991) requires iterative re-training of the classifier and replacing the least relevant feature in the set with a randomly drawn new one.

When the linear network is used for digit classification without frame by frame target labeling, temporal integration of features is necessary. This is done by simple summation of the feature vectors over the whole utterance yielding one feature vector per utterance as required for the linear net. The FFNN approach has been successfully applied to isolated digit recognition with the sigma-pi type of secondary features (Gramß and Strube, 1990, Chapter 3) and also in combination with Gabor features (Chapter 6).

Optimization was carried out on German and English digit targets (zifkom and TIDigits corpora, Leonard, 1984), which are comprised of mainly mono-syllabic words, as well as on parts of the TIMIT corpus (Garofolo, 1998) with phone-based labeling on a frame by frame basis. The phone labels were grouped into a smaller number of classes based on different phonetic features (place and manner of articulation as described by Chang et al. 2001a; Dupont et al. 1997) or, alternatively, only members of a certain single phonetic class (e.g. vowels) were used in the optimization. In addition, optimization experiments were carried out with diphone targets, focusing on the transient elements by using only a context of 30ms to each side of the phoneme boundary. Again, target labels were combined to make the experiments feasible. More than 100 optimization runs were carried out on different data and with different target sets, each resulting in an optimized set of between 10 and 80 features. Apart from the free parameters f_0, ω_f and ω_t the filter mode (real, imaginary or complex) and filter type (spectral only, temporal only, spectro-temporal up, spectro-temporal down) were also varied and equally likely when a new feature is randomly drawn.

The complex filter function (47.7% of all selected features) was consistently preferred over using the real or imaginary part only (cf. Figure A.1 on page 164). This trend is most dominant for ST or purely temporal features, while for spectral features all modes are equally frequent. As can be seen in Figures 7.3 a-f, spectro-temporal (ST) features were selected in 32.7% of all cases. Only minor differences are found on average between using clean or noisy data for the optimization, but significant differences can be

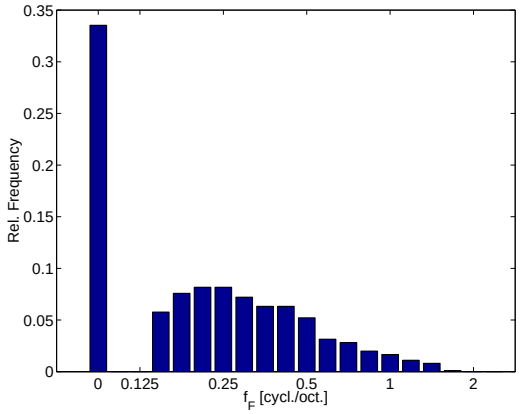


Figure 7.5: Distribution of spectral modulation frequency $\omega_f/2\pi$ for all targets. Purely temporal features accumulate in the 0 bin, although they also have a limited spectral extend.

observed depending on the classification targets. ST features account for 39% of all features in the selected sets for digit targets, while the numbers for diphone and phoneme targets are 33% and 21%, respectively.

There is a significant difference between the phone targets which are grouped according to the manner of articulation with necessary intergroup discrimination and those where only targets of one phonetic class were to be classified (cf. Figures 7.3 e and f). In the former case, ST features were selected less often (9%), while in the latter 28% of all features were ST, with the highest number for diphthongs (46%) and the lowest for stops (14%). For vowels, spectral features dominated (56%) while for stops and nasals the percentage of temporal Gabor functions was highest (41% in both cases, see Figure A.2 on page 165). The feature distribution along the parameter axis of temporal and spectral modulation is plotted in Figures 7.4 and 7.5, respectively. Please note that the parameter values were drawn from a uniform distribution over the $\log_2$ of the modulation frequencies. Temporal modulation frequencies between two to 8Hz dominate with lower modulation frequencies preferred for digit targets and medium (around 8Hz) for diphone targets. Spectral modulation frequencies are consistently preferred to be in the region of 0.2 to 0.7 cycles per octave with only minor differences across target labels. These results correspond well with the importance of different modulation frequencies for speech recognition (Kanedera et al., 1997, 1999), modulation perception thresholds (Chi et al., 1999) and physiological data (Miller et al., 2002).

7.4 ASR Experiments

Recognition experiments were carried out within the Aurora 2 experimental framework (see Hirsch and Pearce, 2000, for details). The fixed aurora HTK back end was trained on multicondition (4 types of noise, 5 SNR levels) or clean only training data. Strings of English digits (from the TIDigits corpus) were then recognized in 50 different noise conditions with 1000 utterances each (10 types of noise and SNR of 0, 5, 10, 15, 20) including convolutive noise. The Tandem recognition system (Hermansky et al., 2000) was used for the Gabor feature sets. Every set of 60 Gabor features is online normalized and combined with delta and double-delta derivatives before feeding into the MLP (60, 1000 and 56 neurons in input, hidden and output layer, respectively), which was trained on the TIMIT phone-labeled database with artificially added noise. The 56 output values are then decorrelated via principal component analysis (PCA, statistics derived on clean TIMIT) and fed into the HTK back end.

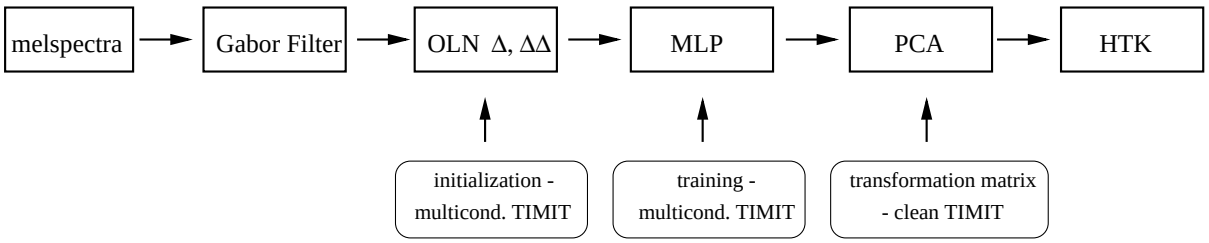


Figure 7.6: Sketch of the Gabor Tandem recognition system as it was used in experiment a).

Figure 7.6 gives a schematic overview of the Gabor/Tandem system. A much better performance can be obtained by carrying out the MLP training on the same corpus as training and testing of the HMM back end. However, in order to obtain a more general front end which is also suitable for other corpora without parameter changes, a different data set (TIMIT) is used here for the derivation of parameters for MLP and PCA. The MLP in the Tandem system performs a non-linear mapping of the input feature set to phoneme posterior probabilities. The output layer non-linearity is given by the softmax function. Each neuron output p_i is calculated from all the summed inputs q_j to J neurons via the following softmax activation function:

$$p_i = \frac{\exp[q_i]}{\sum_{j=1}^{J} \exp[q_j]}. \tag{7.2}$$

To obtain more Gaussian-like distributed features the final non-linearity in the MLP is omitted in forward passing during the test phase (Benitez et al., 2001). The resulting values are quasi log-probabilities which allows for easy combination of two streams by adding the log-posteriors (condition 'P' in Figure 7.7 and Table 7.1). The subsequent PCA decorrelates the features in order to better match the diagonal covariance assumption of the HTK back end. Also, a reduction of feature vector dimension is possible by PCA. This is especially important when two streams are combined by concatenation of the feature vectors (condition 'D', 'Q' in Figure 7.7 and Table 7.1). Diagonal Gabor functions are more frequent in sets which are optimized on diphone target labels. In system G1D, such a set of Gabor features is fed into a MLP trained on diphone labeled data. The resulting features are then concatenated to a phone based Gabor Tandem stream. The Gabor filter sets are shown in detail in Appendix A.

The results in Table 7.1 show a drastic improvement of performance over the reference system (R0) by using the Tandem system, which is further increased by applying Gabor feature extraction (G1, G2, G3) instead of simply using mel-spectra (R1) or mel-cepstra (not shown). Even better performance is obtained by combining Gabor feature streams with mel-spectrum based feature streams via posterior combination (Ellis, 2000, G1P, G3P). Alternatively, improvement may be obtained by concatenation of a Gabor stream with another, diphone-based Gabor stream (G1D) or with the reference stream (G1Q). In all cases the combination of a Gabor feature stream with a non-Gabor stream yields better performance than combining two non-Gabor streams (cf. Figure 7.7). The Tables A.5-A.15 in Appendix A give more detailed results for the Gabor feature sets.

7.5 Summary

An efficient method of feature selection is applied to optimize a set of Gabor filter functions. The underlying distribution of importance of spectral and temporal modulation frequency reflects the properties of speech and is in accordance with physiological and psychoacoustical data. The optimized sets increase the robustness of the Tandem digit recognition system

TIDigits Aurora 2 System description	average WER [%]		WER relative improvement [%]	
	Multi	Clean	Multi	Clean
R0: Aurora2 reference baseline	12.97	41.94	0.00	0.00
R1: Melspec Tandem	12.04	28.66	12.87	40.09
G1: Gabor phone optimized (inter group)	11.68	30.17	14.52	37.19
G2: Gabor phone optimized (inter&intra group)	11.99	26.51	8.40	44.24
G3: Gabor word optimized	11.99	23.63	4.03	51.24
R1-D: melspec phone & diphone	12.86	32.48	8.97	32.38
G1-D: Gabor optimized phone & diphone	11.17	25.29	19.74	50.57
cep-R1-P: posterior combination R1 + mel cepstra	13.45	33.20	13.91	45.08
G1-R1-P: posterior combination Gabor G1 + R1	10.74	24.78	24.64	51.88
G3-R1-P: posterior combination Gabor G3 + R1	10.62	24.73	23.11	53.06
R1-R0-Q: : concatenate R0 & R1	10.74	29.06	25.50	41.98
G1-R0-Q: concatenate R0 & Gabor G1	10.35	27.89	30.45	48.39

Table 7.1: Average word error rate (WER) in percent and average WER reduction relative to the Aurora 2 baseline features (R0). WER and WER reduction are averaged separately over all test conditions. Non-Gabor reference system have gray shading. P denotes posterior combination of two Tandem streams before the final PCA. D indicates the concatenation of two Tandem streams which are optimized on phone and diphone targets, respectively, after reducing the dimension of each to 30 via PCA. Q indicates concatenation of R0 (42 mfcc features) with 18 Tandem features. R1 denotes the Tandem reference system with MLP trained on mel-spectra features in 90ms of context. Gabor set G1 was optimized on noisy TIMIT with broad phonetic classes, G2 on noisy TIMIT for phonetic inter/intra-class discrimination and G3 on noisy German digits (zifkom).

on the TIDigits corpus. This is especially the case when several streams are combined by posterior combination or concatenation, which indicates that the new Gabor features carry complementary information to that of standard front ends.

A major part of this work was carried out at the International Computer Science Institute in Berkeley, California. Special thanks go to Nelson Morgan, Birger Kollmeier, Steven Greenberg, Hynek Hermansky, David Gelbart, Barry Yue Chen, and Stéphane Dupont for their support and

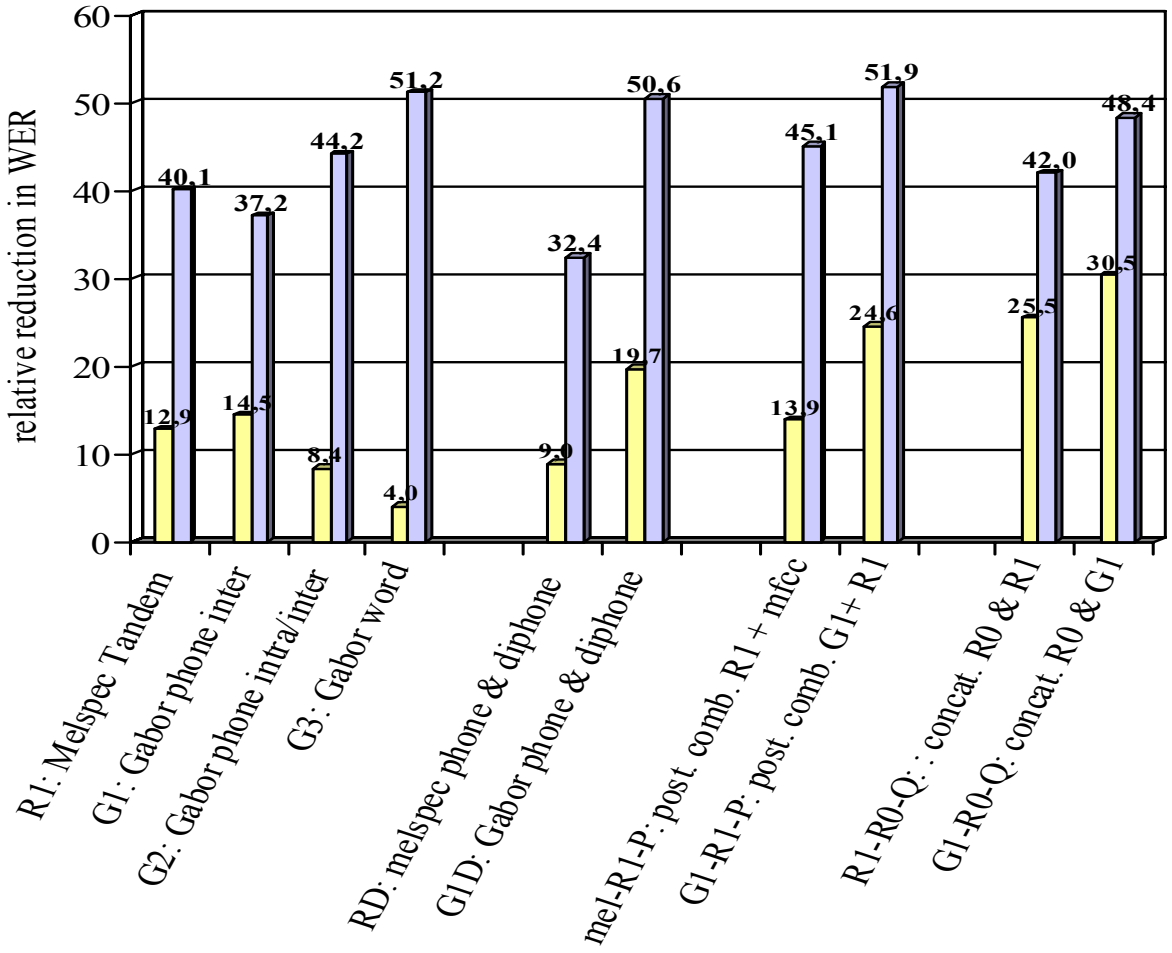


Figure 7.7: Average reduction in word error rate (WER) relative to the Aurora 2 baseline features (R0). Results for clean training (light bars) and multicondition training (dark bars). P denotes posterior combination of two Tandem streams before the final PCA. D indicates the concatenation of two Tandem streams which are optimized on phone and diphone targets, respectively, after reducing the dimension of each to 30 via PCA. Q indicates concatenation of R0 (42 mfcc features) with 18 Tandem features. R1 denotes the Tandem reference system with MLP trained on mel-spectra features in 90ms of context. Gabor set G1 was optimized on noisy TIMIT with broad phonetic classes, G2 on noisy TIMIT for phonetic inter/intra-class discrimination and G3 on noisy German digits (zifkom).

many enlightening discussions. This work was supported by Deutsche Forschungsgemeinschaft (KO 942/15).

chapter 8

Improving Word Accuracy with Gabor Feature Extraction [a]

Contents

8.1 Introduction 138
8.2 Spectro-temporal Feature Extraction 139
8.3 ASR Experiments 141
8.4 Conclusion 147

Abstract

A novel type of feature extraction for automatic speech recognition is investigated. Two-dimensional Gabor functions, with varying extents and tuned to different rates and directions of spectro-temporal modulation, are applied as filters to a spectro-temporal representation provided by mel spectra. The use of these functions is motivated by findings in neurophysiology and psychoacoustics. Data-driven parameter selection was used to obtain optimized Gabor feature sets, the performance of which is evaluated on the Aurora 2 and 3 datasets both on their own and in combination with the Qualcomm-OGI-ICSI Aurora proposal. The Gabor features consistently provide performance improvements. The combination of Qualcomm-OGI-ICSI proposal and Gabor Tandem features yields an average word error rate reduction of 57.0% compared to 50.3% for the Qualcomm-OGI-ICSI proposal alone and 51.1% for the advanced ETSI front end standard.

[a] A slightly modified version of this chapter was published in the *Proceedings of International Conference on Speech and Language Processing (ICSLP) 2002* by Michael Kleinschmidt and David Gelbart.

Zusammenfassung

Ein neuer Typ von Merkmalen zur automatischen Spracherkennung wird untersucht. Zweidimensionale Gabor Funktionen mit variabler Ausdehnung, Ausrichtung und Frequenz werden als spektro-temporale Modulationsfilter eingesetzt. Basis dafür sind Mel-Spektrogramme. Die Verwendung dieser Filter ist motiviert durch Erkenntnisse der neurophysiologischen und psychoakustischen Forschung. Eine datenbasierte Optimierung der Parameter wurde durchgeführt, um optimierte Sätze von Gaborfilter zu erhalten. Diese wurden im Rahmen des Tandem-Erkenners auf den Datensätzen Aurora 2 und 3 evaluiert. Dabei wurde das Gabor-basierte Tandem System sowohl allein als auch in Kombination mit dem Vorschlag von Qualcomm-OGI-ICSI für den ETSI Aurora Standard verwendet. Die Gabor-basierten Merkmale zeigen konsistent eine Verbesserung der Erkennungsleistung. Die Kombination des Qualcomm-OGI-ICSI Systems mit dem Gabor Tandem erreicht eine durchschnittliche Verringerung der Fehlerrate um 57% verglichen mit 50.3% des Qualcomm-OGI-ICSI Systems allein und 51.1% für den neuen ETSI Standard.

8.1 Introduction

Speech is characterized by its fluctuations across time and frequency. The latter reflect the characteristics of the human vocal cords and tract and are commonly exploited in automatic speech recognition (ASR) by using short-term spectral representations such as cepstral coefficients. The temporal properties of speech are targeted in ASR by dynamic (delta and delta-delta) features as well as temporal filtering and feature extraction techniques like RASTA and TRAPS (Hermansky, 1998). Nevertheless, speech clearly exhibits combined *spectro-temporal* modulations. This is due to intonation, coarticulation and the succession of several phonetic elements, e.g., in a syllable. Formant transitions, for example, result in diagonal features in a spectrogram representation of speech. This kind of pattern is explicitly targeted by the feature extraction method used in this paper.

Recent findings from a number of physiological experiments with different mammal species showed that a large percentage of neurons in the primary auditory cortex respond differently to upward versus downward-moving ripples in the spectrogram of the input (Depireux et al., 2001). Each individual neuron is tuned to a specific combination of spectral and temporal

modulation frequencies, with a spectro-temporal response field that may span up to a few 100ms in time and several critical bands in frequency and may have multiple peaks (Schreiner et al., 2000; deCharms et al., 1998). A psychoacoustical model of modulation perception (Chi et al., 1999) was built based on that observation and inspired the use of two-dimensional Gabor functions as a feature extraction method for ASR in this study. Gabor functions are localized sinusoids known to model the characteristics of neurons in the visual system (De-Valois and De-Valois, 1990). The use of Gabor features for ASR has been proposed earlier and proven to be relatively robust in combination with a simple classifier (Chapter 6). Automatic feature selection methods are described in Chapter 7 and the resulting parameter distribution has been shown to remarkebly resemble neurophysiological and psychoacoustical data as well as modulation properties of speech. Other approaches to targeting spectro-temporal variability in feature extraction include time-frequency filtering ('tiffing', Nadeu et al., 2001). Still, this novel approach of spectro-temporal processing by using localized sinusoids most closely matches the neurobiological data and also incorporates other features as special cases: purely spectral Gabor functions perform sub-band cepstral analysis – modulo the windowing function – and purely temporal ones can resemble TRAPS or the RASTA impulse response and its derivatives (Hermansky, 1998) in terms of temporal extent and filter shape.

8.2 Spectro-temporal Feature Extraction

A spectro-temporal representation of the input signal is processed by a number of Gabor functions used as 2-D filters. The filtering is performed by correlation over time of each input frequency channel with the corresponding part of the Gabor function (with the Gabor function centered on the current frame and desired frequency channel) and a subsequent summation over frequency. This yields one output value per frame per Gabor function (these output values are called Gabor features) and is equivalent to a 2-D correlation of the input representation with the complete filter function and a subsequent selection of the desired frequency channel of the output.

In this study, log mel-spectrograms serve as input features for Gabor feature extraction. This representation was chosen for its widespread use in ASR and because the logarithmic compression and mel-frequency scale

might be considered a very simple model of peripheral auditory processing. Any other spectro-temporal representation of speech could be used instead and especially more sophisticated auditory models might be a good alternative in future experiments.

The two-dimensional complex Gabor function $g(t, f)$ is defined as the product of a Gaussian envelope $n(t, f)$ and the complex Euler function $e(t, f)$. The envelope width is defined by standard deviation values σ_f and σ_t, while the periodicity is defined by the radian frequencies ω_f and ω_t with f and t denoting the frequency and time axis, respectively. The two independent parameters ω_f and ω_t allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including *diagonal* modulations. Further parameters are the centers of mass of the envelope in time and frequency t_0 and f_0. In this notation the Gaussian envelope $n(t, f)$ is defined as

$$n(\cdot) = \frac{1}{2\pi\sigma_f\sigma_t} \cdot \exp\left[\frac{-(f-f_0)^2}{2\sigma_f^2} + \frac{-(t-t_0)^2}{2\sigma_t^2}\right] \tag{8.1}$$

and the complex Euler function $e(t, f)$ as

$$e(\cdot) = \exp\left[i\omega_f(f-f_0) + i\omega_t(t-t_0)\right]. \tag{8.2}$$

It is reasonable to set the envelope width depending on the modulation frequencies ω_f and ω_t to keep the same number of periods T in the filter function for all frequencies. Here, the spread of the Gaussian envelope in dimension x was set to $\sigma_x = \frac{\Pi}{\omega_x} = T_x/2$. The infinite support of the Gaussian envelope is cut off at between σ_x and $2\sigma_x$ from the center. For time dependent features, t_0 is set to the current frame, leaving f_0, ω_f and ω_t as free parameters. From the complex results of the filter operation, real-valued features may be obtained by using the real or imaginary part only. In this case, the overall DC bias was removed from the template. The magnitude of the complex output can also be used. Special cases are temporal filters ($\omega_f = 0$) and spectral filters ($\omega_t = 0$). In these cases, σ_x replaces $\omega_x = 0$ as a free parameter, denoting the extent of the filter, perpendicular to its direction of modulation.

8.3 ASR Experiments

8.3.1 Setup

The Gabor features approach is evaluated within the aurora experimental framework (Hirsch and Pearce, 2000) using a) the Tandem recognition system proposed by Hermansky et al. (2000) and d) a combination of it with the Qualcomm-ICSI-OGI proposal for Aurora 2, which is described by Adami et al. (2002). Variants of that are b) and c): the Gabor Tandem system as a single stream combined with noise robustness techniques taken from the Qualcomm-ICSI-OGI proposal.

In all cases the Gabor features are derived from log mel-spectrograms, calculated as specified in ETSI ES 201 v1.1.2 (2000) but modified to output mel-spectra instead of MFCCs, omitting the final DCT. The log mel-spectrogram calculation consists of DC removal, pre-emphasis, Hanning windowing with 10ms offset and 25ms length, FFT and summation of the magnitude values into 23 mel-frequency channels with center frequencies from 124 to 3657Hz. The amplitude values are then compressed by the natural logarithm.

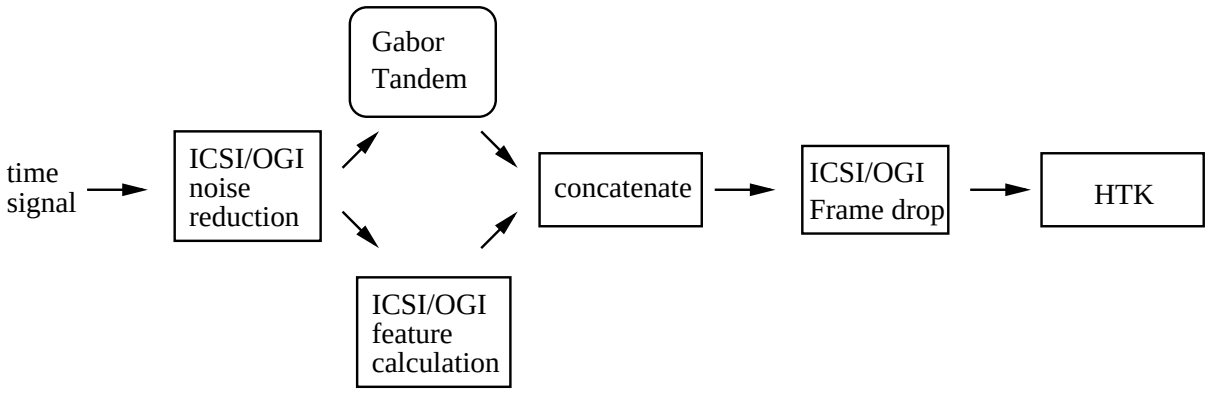


Figure 8.1: Experiment d): Combination of Gabor feature extraction and the Qualcomm-ICSI-OGI proposal system.

Figure 7.6 on page 132 sketches the Tandem system as it is used in experiment a): 60 Gabor filters are fed into a multi-layer perceptron (MLP) after online normalization (OLN) and $\Delta, \Delta\Delta$ processing. The MLP (180 input, 1000 hidden, 56 output units) has been trained on the frame labeled noisy TIMIT corpus (Garofolo, 1998) using frame by frame phoneme targets. The output layer's softmax non-linearity is omitted in forward passing. The resulting 56-dimensional feature vector is then decorrelated

by a PCA transform based on clean TIMIT. The resulting feature vectors are then passed on to the fixed Aurora HTK back end.

Experiment d) is depicted in Figure 8.1. After the initial noise reduction (NR), which is the same as proposed by Adami et al. (2002), a Gabor feature stream identical to that in a) is run in parallel with the Qualcomm-ICSI-OGI proposal feature extraction. The two streams are combined by concatenation before the final dropping (FD) of those frames judged to be nonspeech. The 45 Qualcomm-ICSI-OGI features are combined with a reduced set of 15 features from the Gabor stream which are obtained by reducing the dimension in the PCA stage from 56 to 15. In a variation of this (experiment c), the full set of 56 features from the Gabor stream is used with noise reduction and frame dropping but without concatenating the Qualcomm-ICSI-OGI feature stream. Experiment b) also leaves out the frame dropping stage.

Reference systems are the aurora baseline (R0) front end of 13 mel-cepstral coefficients and their delta and double-deltas used in the unquantized, end-pointed version (ICSLP, 2002), the Qualcomm-ICSI-OGI proposal system (R1), and a combination of R1 with a melspec-based Tandem system (R2). The latter is identical to the Gabor-based Tandem system used apart from the input features to the MLP, which are 23 mel-spectra with deltas and double deltas over 90ms (9 frames) of context. Also, the number of hidden units has been reduced to 300 in order to keep the total number of weights constant. Another references is the combination of R1 with a second stream based on a TRAPs Tandem feature extraction (R1e, Adami et al., 2002) and the new ETSI standard and winner of the Aurora 3 competition (R3, ETSI ES 202 050 v0.1.1, 2002). Both, the new standard and the Qualcomm-ICSI-OGI proposal rely on noise reduction techniques similar to the algorithm by Ephraim and Malah (1984) as well as a TRAP-based voice-activity detection and frame dropping.

In the Aurora 2 experiment, the TIDigits English connected digits corpus (Leonard, 1984) is used for training and testing, artificially mixed with noise of varying levels and types. HTK is trained separately with clean and multi-condition training data. Test set A refers to matched noise (in the case of multicondition training), test set B to mismatched noise and test set C to mismatched channel conditions. For Aurora 3 the SpeechDat-car corpora (Moreno et al., 2000) for Finnish, Spanish, German and Danish (ICSLP, 2002) are used for training and testing. The corpora contain digit strings recorded in various car environments. The experi-

Table 8.1: Aurora 2 (TIDigits): Performance of different front ends in terms of WER and WER reduction relative to the baseline system (R0). The Qualcomm-ICSI-OGI submission system (R1) is compared and combined with different Gabor Tandem (T) systems: Gabor set G1 was optimized on TIMIT phoneme inter-group discrimination, G2 on TIMIT phoneme inter- and within-group discrimination and G3 on German digits. NR indicates noise reduction, FD frame dropping. R2 denotes a Tandem system based on melspectra, R3 the new ETSI standard and R1e the combination of R1 with a TRAPs based Tandem stream. a) refers to a single Tandem stream, b) to a Tandem with NR, c) with NR and FD and d) to the combination of R1 and a Tandem stream as in c).

Aurora 2	WER [%]		Rel. impr. [%]	
	multi	clean	multi	clean
R0: Aurora2 reference	12.97	41.94	0.00	0.00
R1: ICSI/OGI	9.09	15.10	26.41	66.53
R1e: R1 + TRAPs stream	8.25	13.68	35.65	70.11
R3: new ETSI standard	8.43	12.99	32.60	70.86
R2a T melspec	12.04	28.66	12.87	40.09
R2d: R1 + T melspec NR FD	9.18	14.01	34.55	72.29
G1a: T Gabor	11.68	30.17	14.52	37.19
G2a: T Gabor	11.99	26.51	8.40	44.42
G3a: T Gabor	11.99	23.63	4.03	51.24
G1b: T Gabor NR	10.33	16.51	19.88	64.64
G1c: T Gabor NR FD	10.42	14.42	25.74	70.86
G1d: R1 + T Gabor NR FD	8.85	13.04	37.84	74.99
G2d: R1 + T Gabor NR FD	8.70	13.30	37.65	73.88
G3d: R1 + T Gabor NR FD	8.60	12.29	36.40	75.23

mental results refer to well-matched (wm), medium-mismatched (mm) and highly-mismatched (hm) conditions which describe the degree of mismatch of noise and microphone location (close-talking versus hands-free) between the training and test sets. mm indicates a mismatch in noise only, while hm indicates mismatch of noise and microphone.

8.3.2 Feature Selection

The parameters of the 60 Gabor filters were chosen by optimization as described in Chapters 6 and 7. A simple linear classifier was used to evaluate the importance of individual features based on their contribution to classification performance. Gabor set G1 is optimized on inter-group discrimination of phoneme targets from the TIMIT corpus combined into broader phonetic categories of place and manner of articulation as proposed by Chang et al. (2001a) and Dupont et al. (1997). Gabor set G2

Table 8.2: Aurora 2 (TIDigits) and Aurora 3 (SpeechDat-car): Performance of different front ends in terms of WER and WER reduction. Abbreviations as in Table 8.1.

	Aurora 2		Aurora 3		overall	
	WER [%]	impr. [%]	WER [%]	impr. [%]	WER [%]	impr. [%]
R0	27.46	0.00	23.48	0.00	25.47	0.00
R1	12.10	46.47	9.43	53.94	10.77	50.21
R1e	10.96	52.88	9.14	55.72	10.05	54.30
R2d	11.60	53.42	9.23	56.73	10.42	55.08
R3	10.71	51.73	9.90	50.34	10.31	51.04
G1d	10.95	56.41	9.20	57.60	10.08	57.01
G2d	11.00	55.77	8.91	58.28	9.96	57.03
G3d	10.44	55.82	8.88	57.44	9.66	56.63

is optimized on inter- and within-group discrimination of broad phonetic classes, also using the TIMIT corpus. G3 is optimized on German digits (zifkom corpus) using word targets. G1, G2 and G3 contain 27, 28, and 48 filters, respectively, with temporal extents longer than 100 ms, although many in G1 are much shorter. Set G1 consists of 35 features with purely spectral modulation, 23 with purely temporal modulation, and two with spectro-temporal modulation. G2 (34/22/4) and G3 (12/18/30) have a larger number of filters with spectro-temporal modulation. In all three cases, most of the features are two-dimensional in extent, simultaneously occupying more than one frequency channel and time frame. Lists of the filter parameters are given Tables A.1, A.2 and A.3. Also given in the Attachment A, are plots of the Gabor filter functions (Figures A.3-A.5).

8.3.3 Results

The results in Tables 8.1–8.2 are given in absolute word error rate (WER=1-Accuracy) and WER improvement relative to the baseline system (R0). The WER as well as the WER reduction values are averaged over a number of different test conditions in accordance with ICSLP (2002), so the average WER improvement cannot directly be calculated from the average WERs.

All systems in configuration a) yield better results on the Aurora 2 task than the reference system R0 (cf. Table 8.1). The three Gabor sets vary in their performance for clean and noisy training conditions. The more spectro-temporal features are present in the set, the better is the perfor-

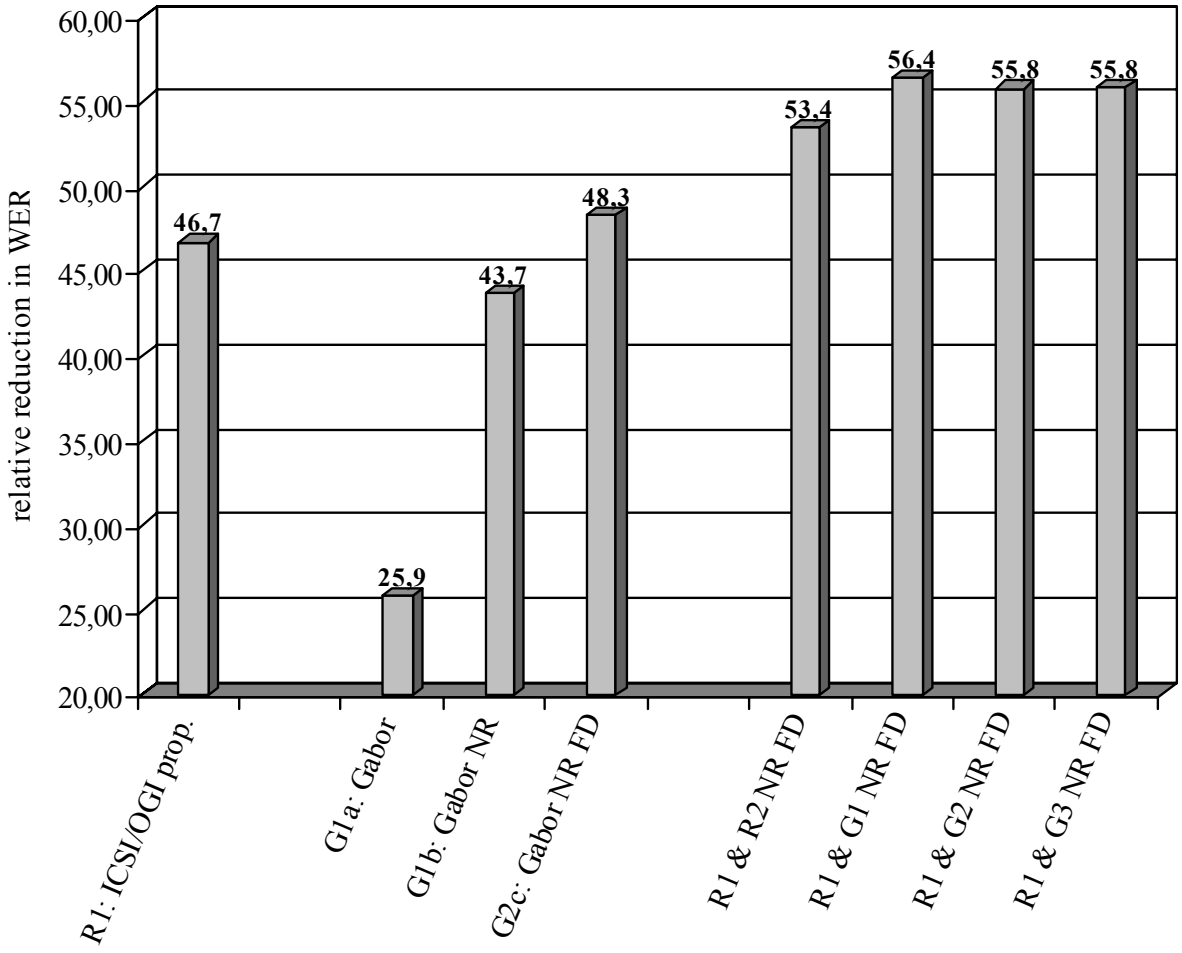


Figure 8.2: Performance on Aurora 2 (TIDigits) in terms of average WER improvement over the baseline (R0). The Qualcomm-ICSI-OGI submission system (R1) is compared and combined with different Gabor Tandem (T) systems: Gabor set G1 was optimized on TIMIT phoneme inter-group discrimination, G2 on TIMIT phoneme inter- and within-group discrimination and G3 on German digits. NR indicates noise reduction, FD frame dropping. R2 denotes a Tandem system based on melspectra. a) refers to a single Tandem stream, b) to a Tandem with NR, c) with NR and FD and d) to the combination of R1 and a Tandem stream as in c).

mance with clean training, indicating an improved robustness with these features. Adding the noise reduction (NR) in b) and the frame dropping stage (FD) in c) further improves the performance.

The best results are obtained by combining the Qualcomm-ICSI-OGI front end R1 with one of the Tandem streams via concatenation in experiment d). Table 8.2 summarizes the results for Aurora 2 and 3. Combining the Qualcomm-ICSI-OGI feature set (R1) with Tandem based features improves performance on Aurora 2 and 3 in terms of average WER and average WER improvement. In terms of average WER improvement the

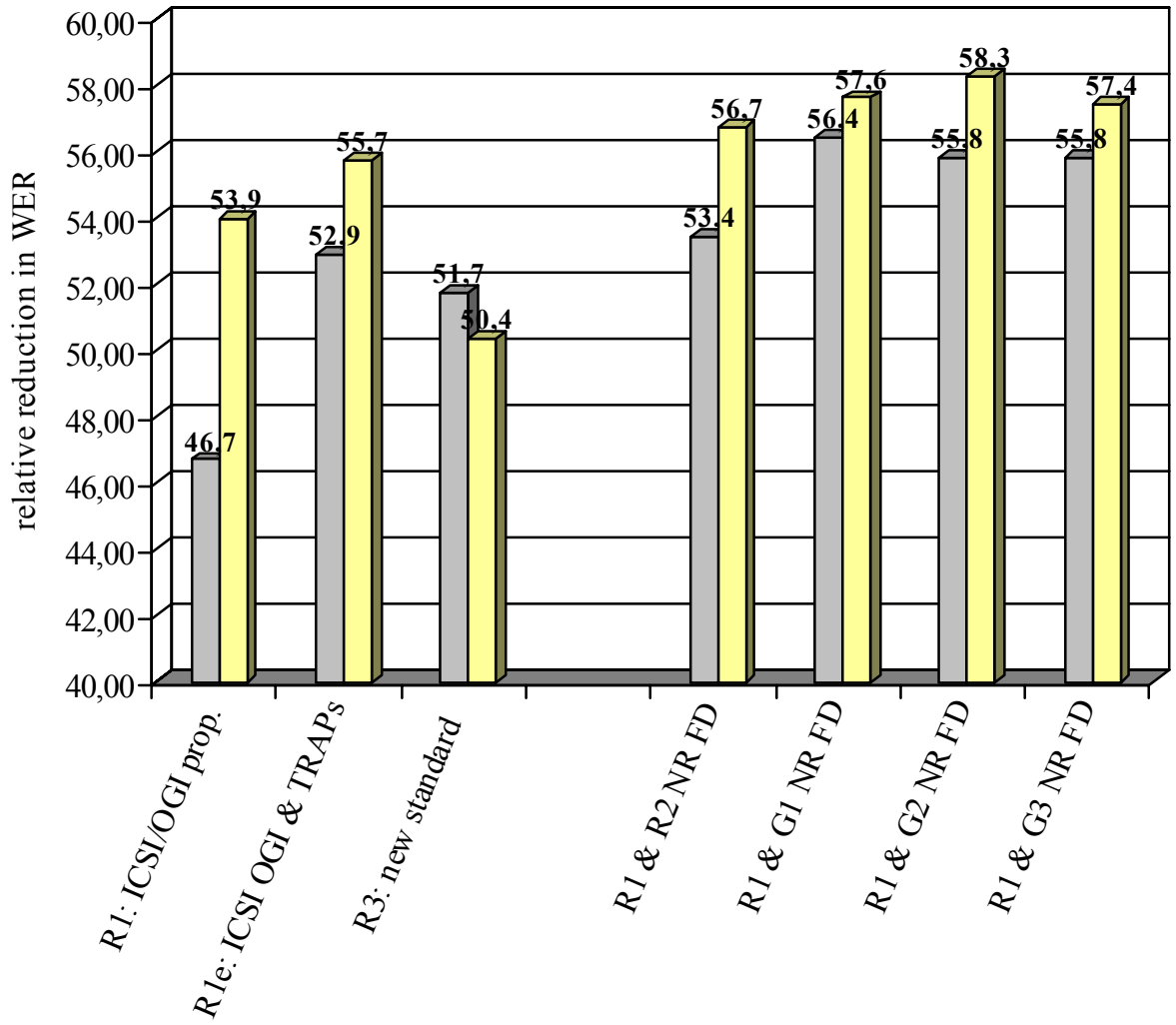


Figure 8.3: Performance on Aurora 2 (TIDigits, dark bars) and 3 (SpeechDat-car, light bars) in terms of average WER improvement over the baseline (R0). The Qualcomm-ICSI-OGI submission system (R1) is compared and combined with different Gabor Tandem (T) systems: Gabor set G1 was optimized on TIMIT phoneme inter-group discrimination, G2 on TIMIT phoneme inter- and within-group discrimination and G3 on German digits. NR indicates noise reduction, FD frame dropping. R2 denotes a Tandem system based on melspectra, R3 the new ETSI standard and R1e the combination of R1 with a TRAPs based Tandem stream. a) refers to a single Tandem stream, b) to a Tandem with NR, c) with NR and FD and d) to the combination of R1 and a Tandem stream as in c).

Gabor Tandem G1d system yields better results than the Qualcomm-ICSI-OGI front end R1 on Aurora 2 (c.f. Figure 8.2). The Gabor based Tandem systems perform better than the mel-spectrogram based Tandem system (R2d). System G2d yields the greatest (57.03%) overall average relative improvement over R0, while system G3d yields the lowest overall WER (9.66%). This is due to G3 being more robust in very adverse conditions, where the absolute gain in WER is higher. As shown in Figure 8.3, the

combination of a Gabor Tandem stream with the Qualcomm-ICSI-OGI feature set (G1d,G2d,G3d) also outperforms the new ETSI standard (R3) and the improved Qualcomm-ICSI-OGI front end with additional TRAPs Tandem stream (R1e). The Tables A.17-A.30 in Appendix A give more detailed results for the Gabor feature sets.

8.4 Conclusion

It has been shown that optimized sets of Gabor features improve robustness when used as part of the Tandem system. When incorporating the Tandem system as a second stream into the already robust Qualcomm-ICSI-OGI proposal, the overall performance can be increased further by almost 7% absolute in average relative WER improvement or over 1% absolute reduction in average WER. The fact that Gabor-based Tandem systems consistently outperformed mel spectrum-based systems shows the usefulness of explicitly targeting extended spectro-temporal patterns. In adverse conditions, the Gabor set G3 with 50% diagonal features performs best, which further supports the approach of spectro-temporal modulation filters. It has to be investigated whether this holds for large vocabulary tasks.

Special thanks go to Barry Yue Chen, Stéphan Dupont, Steven Greenberg, Hynek Hermansky, Birger Kollmeier, Nelson Morgan, and Sunil Sivadas for technical support and great advice.

This work was supported by Deutsche Forschungsgemeinschaft (KO 942/15), the Natural Sciences and Engineering Research Council of Canada, and the German Ministry for Education and Research.

CHAPTER 9

Summary and Conclusions

This thesis documents the development of a new type of feature extraction, based on auditory models, for general signal classification with a special focus on automatic speech recognition (ASR). After starting from an auditory model used as a replacement for standard mel-cepstrum features for ASR, the extraction of secondary features is investigated. Sigma-pi cells and later the more robust Gabor features integrate information over a large region of the spectro-temporal representation consisting of primary feature vectors. This approach allows for explicitly targeting spectro-temporal envelope fluctuations on the feature level and is backed by results of psychoacoustical and neurophysiological studies. Major issues are the automated selection of an optimized set of secondary features and the generalization of this approach from isolated word recognition to word sequences and sub-word units such as phonemes and diphones.

2 In **Chapter 2**, the model of auditory perception (PEMO) after Dau et al. (1996a) is evaluated as a front end for ASR with different back end classifiers. The combination of PEMO and a locally-recurrent neural network (LRNN) performs best. The word error rate (WER) can be further reduced by applying monaural Ephraim and Malah (1984) speech enhancement or binaural filtering (Wittkop et al., 1997) prior to feature extraction. Both noise reduction techniques yield a gain in performance of up to 60% absolute or up to 10dB equivalent SNR improvement on the 10% WER level. The gain is less with non-stationary noise signals for the monaural algorithm and less in reverberant conditions or limited spatial separation of the sources for the binaural algorithm. Nevertheless, performance is never degraded by the pre-processing which also holds for clean test data.

3 The auditory approach is extended to secondary feature extraction for spectro-temporal processing in **Chapter 3**. Sigma-pi cells are

used to derive secondary features based on the PEMO output. Within the Feature-finding Neural Network (FFNN) framework a suitable set of sigma-pi cells is automatically selected by applying the 'substitution rule' (Gramß and Strube, 1990; Gramß, 1991, 1992). The PEMO/FFNN system is found to be superior to the PEMO/LRNN system (which was evaluated in Chapter 2) for isolated digit recognition under a number of noise conditions. Constraining the parameter combinations for the sigma-pi cells to purely spectral or purely temporal processing deteriorates performance. This indicates the importance of 'diagonal' window combinations and therefore of integrated spectro-temporal processing.

4 In **Chapter 4** the sigma-pi approach is developed from isolated word recognition to phoneme recognition, given context and segmentation. The temporal integration has to be limited to a certain number of frames, which adds another parameter to the sigma-pi cells. It is shown that phoneme discrimination is possible with the combination of PEMO primary features and sigma-pi secondary features. The statistics of the feature parameters in the optimized sets reflect the class of phonemes to be discriminated (e.g. small temporal extension for vowels, spectro-temporal features for diphthongs). This should be regarded as a proof of concept for classification of small sub-word units within the auditory framework. It is a prerequisite for later application in continuous recognition tasks (Chapters 7 and 8).

5 The problem of long-term sub-band SNR estimation is addressed in **Chapter 5**. PEMO/sigma-pi cell features are combined with the FFNN linear network for feature set optimization and a non-linear neural network for continuous SNR estimation. An estimation error of 5.68dB SNR is reached on speech mixed with previously unknown realistic noise data. The real-world recordings include non-stationary and even speech-like modulated signals such as alarm sounds. Detailed analysis reveals that temporal modulation of three to 11Hz in the noise signal leads to an overestimation the SNR. The performance is comparable to that of algorithms known from the literature, even though only low-frequency spectro-temporal modulations are used in this approach. The estimation error increases by more than 1dB if no spectro-temporal window combinations are allowed, which confirms the importance of spectro-temporal processing found in Chapters 3 and 4. At first sight, the successful application of the PEMO/sigma-pi approach to SNR estimation seems to contradict the demands of a robust front end for ASR, which should be invariant under different noise conditions. Normal hearing human listeners easily localize

and identify other non-speech sound sources, which is indeed a prerequisite for successful auditory scene analysis. Therefore, the auditory approach should not only be useful for ASR, but also for other signal classification tasks such as SNR estimation. It is important to keep this in mind when designing an auditory model based front end. In the case of the sigma-pi approach to ASR and SNR estimation this is taken care of by individually optimizing the feature sets for both applications.

6 The sigma-pi cell method for secondary feature extraction is extended in **Chapter 6** to multiple-window cells and stochastic combination of windows. In addition, Gabor filtering is introduced as a new method for capturing spectro-temporal modulations. In contrast to sigma-pi cells, Gabor filtering is a linear operation. While the stochastic combination of windows allows for "OR" as well as for "AND" operations between windows, its performance on clean test data shows word error rates of over three percent compared to about one percent for the other techniques. In terms of robustness in additive noise, all types of secondary features show good and comparable performance in isolated digit recognition experiments. It is remarkable that the Gabor features in combination with the FFNN back end outperform the combination of Aurora baseline feature and Hidden Markov Model (HMM) in some conditions, even though the FFNN system simply uses a linear classifier.

7 The Gabor filtering approach is further explored in **Chapter 7**. Mel-spectrogram based Gabor features are optimized using the FFNN system. However, the Tandem recognition system (Hermansky et al., 2000) is applied in ASR experiments with the Aurora 2 experimental framework for small vocabularies (Hirsch and Pearce, 2000), instead of using the linear classifier of the FFNN. The feature selection is carried out on different corpora and target labels: digits (mainly mono-syllabic), diphones and phonemes. Diagonal features are very important for digits and diphone optimization, where those chirp-like spectro-temporal patterns are selected more frequently than for phoneme labels. The overall distribution of temporal and spectral modulation frequencies in the sets reflects the properties of speech (e.g. peaks at 4Hz) and is in accordance with neurophysiological and psychoacoustical data. The Gabor Tandem system yields a reduction in word error rate (WER) of up to 50% relative for clean condition training, whether alone or in combination with a second, conventional stream, and 20 to 30% relative for multi-condition training when combined with a second, conventional stream. Any tested combination of a Gabor stream with a another stream yields significantly lower WER than a

comparable combination of two conventional streams. This highlights the importance of the Gabor feature set and the complementary information it carries when compared to conventional streams.

8 In **Chapter 8** the ASR experiments described in Chapter 7 are carried out with additional noise reduction techniques (Wiener filtering and frame dropping). In terms of relative improvement over the baseline system for Aurora 2, the Gabor Tandem with noise reduction (48.3%) is superior to the Qualcomm-ICSI-OGI system (46.7%). The experiments are also extended to the Aurora 3 task, which includes the multi-lingual SpeechDat-car corpus (Finnish, Spanish, German and Danish). The Gabor Tandem features are evaluated with the additional noise reduction techniques and in combination with the Qualcomm-ICSI-OGI proposal as a second stream. When the two streams are combined, an average reduction in word error rate of 57.0% is obtained on the Aurora 2 and 3 tasks. This outperforms the Qualcomm-ICSI-OGI alone (50.3%), the combination of Qualcomm-ICSI-OGI and a TRAPs stream (54.3%) and also the new standard (51.1%) and winner of the Aurora 3 competiton (ETSI ES 202 050 v0.1.1, 2002).

Within the ETSI aurora competition the aim was to develop a new standard front end for distributed ASR which is to be incorporated into mobile phones. Therefore, certain restrictions applied to the feature extraction algorithms which were submitted. For example, the overall algorithmic latency had to be below 240ms and there were also constraints concerning the computational complexity. At least the former restriction was violated by the sets of Gabor features used in Chapters 7 and 8, as some of the Gabor filters had a temporal extend of more than 240ms backward and forward in time. Nevertheless it is remarkable that the new spectro-temporal Gabor features gained an increase in performance compared to state-of-the-art front ends, which have been specifically designed to become the new ETSI standard in a highly demanding selection process. This could only be achieved due to the Tandem system, which consists of a neural network based non-linear mapping of the input features into phoneme log-posterior probabilities. After decorrelation via PCA the resulting features fit very well to an HMM back end, even when its configuration parameters are fixed as in the Aurora framework. Normally, the interaction of front end and classifier requires a large amount of parameter tuning for novel features to become competitive (if at all). This problem is often encountered and

specifically addressed by Bourlard et al. (1996b) in their plea for new ideas "Towards *increasing* speech recognition error rate".

It is also worth noting that with the new ETSI standard for distributed ASR front ends (ETSI ES 202 050 v0.1.1, 2002), noise reduction techniques like Wiener filtering and Ephraim-Malah algorithm as well as the frame dropping stage become the quasi-standard for ASR front ends. The former is also proposed in Chapter 2 and the latter is based on voice activity detection which is related to the SNR estimation problem addressed in Chapter 5.

In this thesis, the new approach of spectro-temporal feature extraction for signal processing and especially ASR applications, is thoroughly investigated. It is shown many times that the restriction to purely spectral or purely temporal processing is not sufficient. Especially the Gabor filter method yields features which increase the performance of state-of-the-art ASR systems. This is only the beginning, as a lot of research and optimization is still to be carried out:

- **large vocabulary:** The new type of spectro-temporal feature extraction has to be tested on large vocabulary tasks, a major goal of ASR. This requires the classification of sub-word units such as phonemes, diphones or syllables. The feasibility of the sigma-pi cell features for phoneme classification is investigated in Chapter 4 and the Tandem network in Chapter 7 is trained on all English phonemes already, rather than on only those which occur in the digits, allowing for an easy change to other languages as it can be seen in Chapter 8.

- **auditory primary feature matrix:** The experiments in Chapters 2-5 are already based on the model of auditory perception (Dau et al., 1996a) to produce the primary feature matrix. The experiments in Chapters 7 and 8 are still to be repeated with the PEMO primary features. There are other models which could be investigated, too. Of special interest is the early auditory spectrogram (Wang and Shamma, 1994), which was successfully combined with spectro-temporal processing by Chi et al. (1999) for modulation perception prediction. Both models are somewhat comparable, as for example the same type of peripheral filterbank is used, and differ mainly in the type of contrasting that is applied to enhance transitions. The model of auditory perception mainly enhances temporal transitions using non-linear adaptation loops. In contrast, the early

auditory spectrogram includes a lateral inhibition stage for contrasting between neighboring frequency channels.

- **syllable-based ASR:** It is very likely that the extended spectro-temporal features are even more suitable for syllable classification than for the phoneme-based Tandem processing investigated in Chapters 7 and 8. This is due to the temporal extend of the Gabor features and the good performance on (mainly monosyllabic) digit corpora. It would be beneficial to carry out the feature set optimization on a syllable labeled corpus. The syllable is also the most natural unit of speech, as syllables are relatively easy to segment automatically. Furthermore, the variability of phonemes to a large degree depends on the surrounding syllable (Greenberg, 1999). However, the number of syllables exceeds the number of phonemes by orders of magnitude and a syllable-based front end, although favored, is harder to build due to the computational complexity involved and the large amounts of training data required.

This thesis demonstrates the benefit of auditory modeling in speech processing. Especially when applying a distinct spectro-temporal modulation filtering to the primary feature matrix, in the form of sigma-pi cells or the refined Gabor filters, the performance of ASR systems can be increased significantly. This has strong implication on the field of robust ASR front end development, where the competition of purely spectral and purely temporal feature extraction is still undecided, yet the two types of features are merely special cases of the more general two-dimensional feature types. The results above also show that it is worthwhile to integrate information over time *and* frequency on the feature level and that a parametric filter function such as the Gabor function is suitable for this task. The data-driven automatic feature selection process yields a significant number of diagonal filter shapes. Therefore, it indicates that this type of sensitivity is also reflected in the receptive fields of neurons in the human auditory cortex, which cannot be measured by the type of invasive neurophysiological experiments cited above. It is reasonable to assume that this important type of information is not missed by the human auditory system.

Bibliography

Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., P. Jain, S. K., Morgan, N. and Sivadas, S. (2002). QUALCOMM-ICSI-OGI features for ASR. In *ICSLP*. Denver, Colorado, USA. submitted.

Allen, J. B. (1994). How do humans process and recognize speech. *IEEE Trans. Speech Audio Processing*, **2**(4):567–576.

Avendano, C., Hermansky, H., Vis, M. and Bayya, A. (1996). Adaptive speech enhancement using frequency-specific SNR estimates. In *Proc. IEEE IVTTA'96*, pp. 65–68. Basking Ridge, N.J.

Benitez, C., Burget, L., Chen, B., Dupont, S., Garudadri, H., Hermansky, H., Jain, P., Kajarekar, S. and Sivadas, S. (2001). Robust ASR front-end using spectral-based and discriminant features: experiments on the aurora tasks. In *Proc. Eurospeech 2001, Aalborg*.

Bitzer, J., Simmer, K. and Kammeyer, K.-D. (1999). Multi-microphone reduction techniques for hands-free speech recognition - a comparative study. In *Proc. of Workshop on Robust Methods for Speech Recognition*, pp. 171–174. Tampere, Finland.

Blauert, J. (1997). *Spatial Hearing*. MIT Press, Cambridge, Massachusetts, revised edition.

Bodden, M. and Anderson, T. (1995). Binaurale automatische Spracherkennung im Störschall. In *Fortschritte der Akustik - DAGA 1995*, pp. 1145–1148. DEGA, Oldenburg.

Bourlard, H., Dupont, S., Hermansky, H. and Morgan, N. (1996a). Towards sub-band-based speech recognition. In *European Signal Proc. Conf., Trieste*, pp. 1579–1582.

Bourlard, H., Hermansky, H. and Morgan, N. (1996b). Towards increasing speech recognition error rate. *Speech Communication*, **18**:205–231.

Bourlard, H. and Morgan, N. (1998). Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. In *Adaptive Processing of Sequences and Data Structures*, Vol. 1387 of *Lecture Notes in Artificial Intelligence*, pp. 389–417. Giles, C.L. and Gori, M.

Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organization of sound*. MIT Press, Cambridge, Massachusetts.

Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. on Speech and Audio Proc.*, **2**(2):345–349.

Chang, S., Greenberg, S. and Wester, M. (2001a). An elitist approach to articulatory-acoustic feature classification. In *Proc. Eurospeech*.

Chang, S., Lokendra, S. and Greenberg, S. (2000). Automatic phonetic transcription of spontaneous speech (American English). In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*.

Chang, S., Shastri, L. and Greenberg, S. (2001b). Robust phonetic feature extration under a wide range of noise backgrounds and signal-to-noise ratios. In *Proc. Workshop on consistent and reliable acoustic cues for sound analysis*.

Chi, T., Gao, Y., Guyton, M. C., Ru, P. and Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.*, **106**(5):2719–2732.

Colburn, H. S. (1996). Computational models of binaural processing. In *Auditory Computation* (published by Hawkins, H. L., McMullen, T. A., Popper, A. N. and Fay, R. R.), Springer Handbook of Auditory Research, Chap. 8, pp. 332–400. Springer, New York.

Cooke, M., Green, P., Josifovski, L. and A., V. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, **34**:267–285.

Dau, T., Kollmeier, B. and Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation: I. Modulation detection and masking with narrowband carriers. *J. Acoust. Soc. Am.*, **102**(2):2892–2905.

Dau, T., Kollmeier, B. and Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration. *J. Acoust. Soc. Am.*, **102**:2906–2919.

Dau, T., Püschel, D. and Kohlrausch, A. (1996a). A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure. *J. Acoust. Soc. Am.*, **99**(6):3615–3622.

Dau, T., Püschel, D. and Kohlrausch, A. (1996b). A quantitative model of the 'effective' signal processing in the auditory system: II. Simulations and measurements. *J. Acoust. Soc. Am.*, **99**(6):3623–3631.

De-Valois, R. and De-Valois, K. (1990). *Spatial Vison*. Oxford U.P., New York.

deCharms, R. C., Blake, D. T. and Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, **280**:1439–1443.

Depireux, D., Simon, J., Klein, D. and Shamma, S. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, **85**:1220–1234.

Derleth, R. P. (1999). *Temporal and compressive properties of the normal and impaired auditory system*. Dissertation, Universität Oldenburg.

Dupont, S. and Ris, C. (1999). Assessing local noise level estimation methods. In *Proc. Workshop on robust methods for speech recognition in adverse environments*, pp. 115–118. Tampere, Finland.

Dupont, S. and Ris, C. (2001). Assessing local noise level estimation methods: Application to noise robust ASR. *Speech Communication*, **34**:141–158.

Dupont, S., Ris, C., Deroo, O., Fontaine, V., Boite, J. and Zanoni, L. (1997). Context independent and context dependent hybrid HMM/ANN systems for vocabulary independent tasks. In *Proc. Eurospeech*. ESCA, Rhodes, Greece.

Durlach, N. I. (1972). Binaural signal detection: Equalization and cancellation theory. In *Foundations of modern auditory theory* (published by Tobias, J. V.), Vol. II, Chap. 10, pp. 369–462. Academic Press, New York.

Ellis, D. (2000). Improved recognition by combining different features and different systems. In *Proc. AVIOS*.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, **ASSP-32**(6):1109–1121.

Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log spectral amplitude estimator. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, **ASSP-33**(2):443–445.

ETSI ES 201 v1.1.2 (2000). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms.

ETSI ES 202 050 v0.1.1 (2002). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced feature extraction algorithm.

Fischer, A. and Stahl, V. (1999). On improvement measures for spectral subtraction applied to robust automatic speech recognition in car environments. In *Proc. of Workshop on Robust Methods for Speech Recognition*, pp. 75–78. Tampere, Finland.

Fletcher, H. (1953). *Speech and Hearing in Communication.* Krieger. (There is a 1994 reprint ASA Edition.).

Francis, I. F. and Anderson, T. R. (1997). Binaural phoneme recognition using the auditory image model and cross-correlation. In *Proc. ICASSP 97*, pp. 1231–1234.

Furui, S. (1986). Speaker independend isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoust., Speech, Signal Processing*, **34**:52–59.

Garofolo, J. (1998). Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. National Institute of Standards and Technology (NIST). Gaithersburgh, Maryland.

Gelin, P. and Junqua, J.-C. (1999). Techniques for robust speech recognition in the car environment. In *Proc. Eurospeech 1999*, Vol. 6, pp. 2483–2486. Budapest, Hungary.

Ghitza, O. (1988). Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in noisy environment. *Journal of Phonetics*, **16**:109–123.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, **16**:261–291.

Gramss, T. (1991). Fast algorithms to find invariant features for a word recognizing neural net. In *IEEE 2nd International Conference on Artificial Neural Networks*, pp. 180–184. Bournemouth.

Gramss, T. (1992). *Worterkennung mit einem künstlichen neuronalen Netzwerk*. Dissertation, Universität Göttingen.

Gramss, T. and Strube, H. W. (1990). Recognition of isolated words based on psychoacoustics and neurobiology. *Speech Communication*, **9**:35–40.

Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, **29**:159–176.

Greenberg, S., Arai, T. and Silipo, R. (1998). Speech intelligibility derived from exceedingly sparse spectral information. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*.

Hagan, M. T. and Menhaj, M. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, **5**(6):989–993.

Hansen, M. and Kollmeier, B. (1997). Using a quantitative psychoacoustical signal representation for objective speech quality measurement. In *Proc. ICASSP 1997, Munich*, pp. 1387–1391.

Hansen, M. and Kollmeier, B. (2000). Objective modeling of speech quality with a psychoacoustically validated auditory model. *J. Audio Eng. Soc.*, **48**(5):395–409.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.*, **87(4)**:1738–1752.

Hermansky, H. (1998). Should recognizers have ears? *Speech Communication*, **25**:3–24.

Hermansky, H., Ellis, D. and Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *Proc. ICASSP, Istanbul*.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Trans. Speech Audio Processing*, **2**(4):578–589.

Hermansky, H. and Pavel, M. (1998). RASTA model and forward masking. In *Proc. Computational Hearing*. ASI (Advanced Study Institue), Il Ciocco.

Hermansky, H. and Sharma, S. (1998). TRAPS - Classifiers of temporal patterns. In *Proc. ICSLP'98*, Vol. 3, pp. 1003–1006.

Hermus, K., Dologlou, I., Wambacq, P. and Van Compernolle, D. (1999). Fully adaptive SVD-based noise removal for robust speech recognition. In *Proc. Eurospeech 1999*, Vol. 5, pp. 1951–1954. Budapest, Hungary.

Hirsch, H. and Pearce, D. (2000). The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000, Paris - Automatic Speech Recognition: Challenges for the Next Millennium*.

Hirsch, H. G. (1993). Estimation of noise spectrum and its applications to SNR-estimation and speech enhancement. Technical Report TR-93-012, International Computer Science Institute, Berkeley, California, USA.

Hirsch, H. G. and Ehrlicher, C. (1995). Noise estimation techniques for robust speech recognition. In *Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, pp. 153–156. IEEE.

Hohmann, V. (2002). Gammatone filter bank and re-synthesis. *Acustica united with Acta Acustica*. (accepted for publication).

Holube, I. and Kollmeier, B. (1996). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J. Acoust. Soc. Am.*, **100**:1703–1716.

ICRA (1997). International Collegium of Rehabilitory Audiology - Hearing Aid Clinical Test Environment Standardization Work Group: ICRA noise signals, version 0.3. CDROM.

ICSLP (2002). http://ICSLP2002.colorado.edu/special_sessions/aurora.

Jankowski, C., Hoang-Doan, H. and Lippmann, R. (1995). A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. on Speech and Audio Processing*, **3**(4):286–293.

John, G., Kohavi, R. and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine Learning: Proc. of the 11th international conference*, pp. 121–129. Morgan Kaufmann, San Francisco, CA.

Kaernbach, C. (2000). Early auditory feature coding. In *Contributions to psychological acoustics: Results of the 8th Oldenburg Symposium on Psychological Acoustics.*, pp. 295–307. BIS, Universität Oldenburg.

Kajarekar, S., Yegnanarayana, B. and Hermansky, H. (2001). A study of two dimensional linear discriminants for ASR. In *Proc. ICASSP 2001, Salt Lake City.*

Kanedera, N., Arai, T., Hermansky, H. and Pavel, M. (1997). On the importance of various modulation frequencies for speech recognition. In *Proc. Eurospeech*, pp. 1079–1082. ESCA, Rhodes, Greece.

Kanedera, N., Arai, T., Hermansky, H. and Pavel, M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, **28**:43–55.

Kasper, K. and Reininger, H. (1999). Evaluation of PEMO in robust speech recognition. *J. Acoust. Soc. Am.*, **105**(2):1175.

Kasper, K., Reininger, H. and Wolf, D. (1997). Exploiting the potential of auditory preprocessing for robust speech recognition by locally recurrent neural networks. In *Proc. ICASSP 1997*, pp. 1223–1226.

Kasper, K., Reininger, H., Wolf, D. and Wüst, H. (1995). A speech recognizer with low complexity based on RNN. In *Neural Networks for Signal Processing V, Proc. of the IEEE Workshop, Cambridge (MA)*, pp. 272–281.

Kermorvant, C. and Morris, A. (1999). A comparison of two strategies for ASR in additive noise: Missing data and spectral subtraction. In *Proc. Eurospeech 1999*, Vol. 6, pp. 2841–2844. Budapest, Hungary.

Kingsbury, B., Morgan, N. and Greenberg, S. (1998). Robust speech recognition using the modulation spectrogram. *Speech Communication*, **25**(1):117–132.

Kiyohara, K., Kaneda, Y., Satoshi, T., Hiroaki, N. and Junji, K. (1997). A microphone array system for speech recognition. In *Proc. ICASSP 1997*, pp. 215–218.

Kleinschmidt, M., Marzinzik, M. and Kollmeier, B. (1999). Combining monaural noise reduction algorithms and perceptive preprocessing for robust speech recognition. In *Psychophysics, physiology and models of hearing* (published by Dau, T., Hohmann, V. and Kollmeier, B.), pp. 267–270. World Scientific, Singapore.

Kleinschmidt, M., Tchorz, J., Wittkop, T., Hohmann, V. and Kollmeier, B. (1998). Robuste Spracherkennung durch binaurale Richtungsfilterung und gehörgerechte Vorverarbeitung. In *Fortschritte der Akustik - DAGA 1998*, pp. 396–397. DEGA, Oldenburg.

Kohler, K., Lex, G., Pätzold, M., Scheffers, M., Simpson, A. and Thon, W. (1994). Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL-3.0. Techn. Ber., Verbmobil-Technischer Report.

Kollmeier, B. and Koch, R. (1994). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *J. Acoust. Soc. Am.*, **95**(3):1593–1602.

Kollmeier, B., Peissig, J. and Hohmann, V. (1993). Real-time multiband dynamic compression and noise reduction for binaural hearing aids. *J. Rehab. Res. Dev.*, **30**:82–94.

Kollmeier, B., Sotscheck, J. and Kammermeier (1988). Digitalaufnahme eines Reimtests in deutscher Sprache. *Audiol. Akustik*, **27**:24–27.

Kowalski, N., Depireux, D. and Shamma, S. (1996). Analysis of dynamic spectra in ferret primary auditory cortex. I. Prediction of unit responses to moving ripple spectra. *J. Neurophysiol.*, **76**:3503–3523.

Leonard, R. (1984). A database for speaker independent digit recognition. In *Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, Vol. 3.

Lin, C.-T., Nein, H.-W. and Hwu, J.-Y. (2000). GA-based noisy speech recognition using two-dimensional cepstrum. *IEEE Trans. Speech Audio Processing*, **8**(6):664–675.

Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication*, **22**:1–15.

Martin, R. (1993). An efficient algorithm to estimate the instantaneous SNR of speech signals. In *Proc. Eurospeech*, pp. 1093–1096. ESCA.

Marzinzik, M. and Kollmeier, B. (1999). Developement and evaluation of single-microphone noise reduction algorithms for digital hearing aids. In *Psychophysics, physiology and models of hearing* (published by Dau, T., Hohmann, V. and Kollmeier, B.). World Scientific, Singapore.

Marzinzik, M., Wittkop, T. and Kollmeier, B. (1999). Combination of monaural and binaural noise suppression algorithms and its use for the hearing impaired. *J. Acoust. Soc. Am.*, **105**(2):977.

Meyer, J. and Simmer, K. (1997). Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtracion. In *Proc. ICASSP 1997*, Vol. 2, pp. 1167–1170.

Miller, L., Escabi, M., Read, H. and Schreiner, C. (2002). Spectrotemporal receptive fields in the lemniscal auditory cortex. *Journal of Neurophysiology*, **87**:516–527.

Mine, R., Kobayashi, T. and Katsuhiko, S. (1996). Speech recognition in nonstationary noise based on parallel HMMs and spectral subtraction. *Systems and Computers in Japan*, **27**(14):37–44.

Moore, B. C. J. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.*, **74**:750–753.

Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K. and Allen, J. (2000). Speechdat-car a large speech database for automotive environments. In *LREC (Language Resources and Evaluation), Athens, 2000.*

Müsch, H. and Buus, S. (2001). Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance. *J. Acoust. Soc. Am.*, **109**(6):2910–2920.

Nadeu, C., Macho, D. and Hernando, J. (2001). Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication*, **1–2**:93–114.

Omologo, M., Matassoni, M., Svaizer, P. and Giuliani, D. (1997). Microphone array based speech recognition with different talker-array positions. In *Proc. ICASSP 1997*, pp. 227–230.

Patterson, R. D., Nimmo-Smith, J., Holdsworth, J. and Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. Paper presented at a meeting of the IOC Speech Group on Auditory Modelling at RSRE.

Peissig, J. (1993). *Binaurale Hörgerätestrategien in komplexen Störschallsituationen*, Vol. 88 of *17*. VDI Verlag, Düsseldorf.

Peissig, J. and Kollmeier, B. (1997). Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *J. Acoust. Soc. Am.*, **101**(3):1660–1670.

Püschel, D. (1988). *Prinzipien der zeitlichen Analyse beim Hören*. Dissertation, Universität Göttingen.

Schreiner, C. and Calhoun, B. (1994). Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions. *Auditory Neuroscience*, **1**:39–61.

Schreiner, C., Read, H. and Sutter, M. (2000). Modular organization of frequency integration in primary auditory cortex. *Annual Review Neuroscience*, **23**:501–529.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, **16**:55–76.

Sharma, S. (1999). *Multi-Stream Approach To Robust Speech Recognition*. Dissertation, OGI, Portland, USA.

Siemens (1992). CDROM for fitting and testing of hearing programs.

Somervuo, P. (2002). Experiments with linear and nonlinear feature transformations in HMM based phone recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*. (submitted).

Tchorz, J., Kasper, K., Reininger, H. and Kollmeier, B. (1997). On the interplay between auditory-based features and locally recurrent neural networks. In *Proc. Eurospeech 1997*, Vol. 4, pp. 2075–2078. Rhodes, Greece.

Tchorz, J., Kleinschmidt, M., Kasper, K. and Kollmeier, B. (1999). Auditory feature extraction and recognizer dependencies. In *Proc. of Workshop on Robust Methods for Speech Recognition*, pp. 67–70. Tampere, Finland.

Tchorz, J., Kleinschmidt, M. and Kollmeier, B. (2001). Noise suppression based on neurophysiologically-motivated SNR estimation for robust speech recognition. In *Advances in Neural Information Processing Systems 13 - NIPS 2000* (published by Leen, T. K., Dietterich, T. G. and Tresp, V.), pp. 821–827. MIT Press.

Tchorz, J. and Kollmeier, B. (1999a). A model of auditory perception as front end for automatic speech recognition. *J. Acoust. Soc. Am.*, **106**(4):2040–2050.

Tchorz, J. and Kollmeier, B. (1999b). Speech detection and SNR prediction basing on amplitude modulation pattern recognition. In *Proc. Eurospeech*, pp. 2399 – 2404. ISCA, Budapest, Hungary.

Tchorz, J. and Kollmeier, B. (2001). Estimation of the signal-to-noise ratio with amplitude modulation spectrograms. *Speech Communication*. (accepted for publication).

Varga, A., Steeneken, H., Tomlinson, M. and Jones, D. (1992). The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, UK, & TNO, The Netherlands.

Vizinho, A., Green, P., Cooke, M. and Josifovski, L. (1999). Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study. In *Proc. Eurospeech 1999*, Vol. 5, pp. 2407–2410. Budapest, Hungary.

Wang, K. and Shamma, S. (1994). Self-normalization and noise robustness in early auditory representations. *IEEE Trans. Speech Audio Process.*, **3**:382–395.

Warren, R. and Bashford, J. J. (1999). Intelligibility of 1/3-octave speech: Greater contribution of frequencies outside than inside the nominal. *J. Acoust. Soc. Am.*, **106**(5):L47–L52.

Weber, K., Bengio, S. and Bourlard, H. (2000). HMM2 - A novel approach to HMM emission probability estimation. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*.

Wesselkamp, M. (1994). *Messung und Modellierung der Verständlichkeit von Sprache*. Dissertation, Universität Göttingen.

Wilmers, H. and Strube, H. W. (1999). Noise reduction for speech signals by operations on the modulation frequency spectrum. *J. Acoust. Soc. Am.*, **105**(2):1092.

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W. and Kollmeier, B. (1997). Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction. *Acustica united with acta acustica*, **83**(4):684–699.

Wittkop, T., Hohmann, V. and Kollmeier, B. (1999). Noise reduction strategies employing interaural parameters. *J. Acoust. Soc. Am.*, **105**(2):977.

Zerbs, C. (1999). *Modelling the effective binaural signal processing in the auditory system*. Dissertation, Universität Oldenburg.

APPENDIX A

Tables and Figures

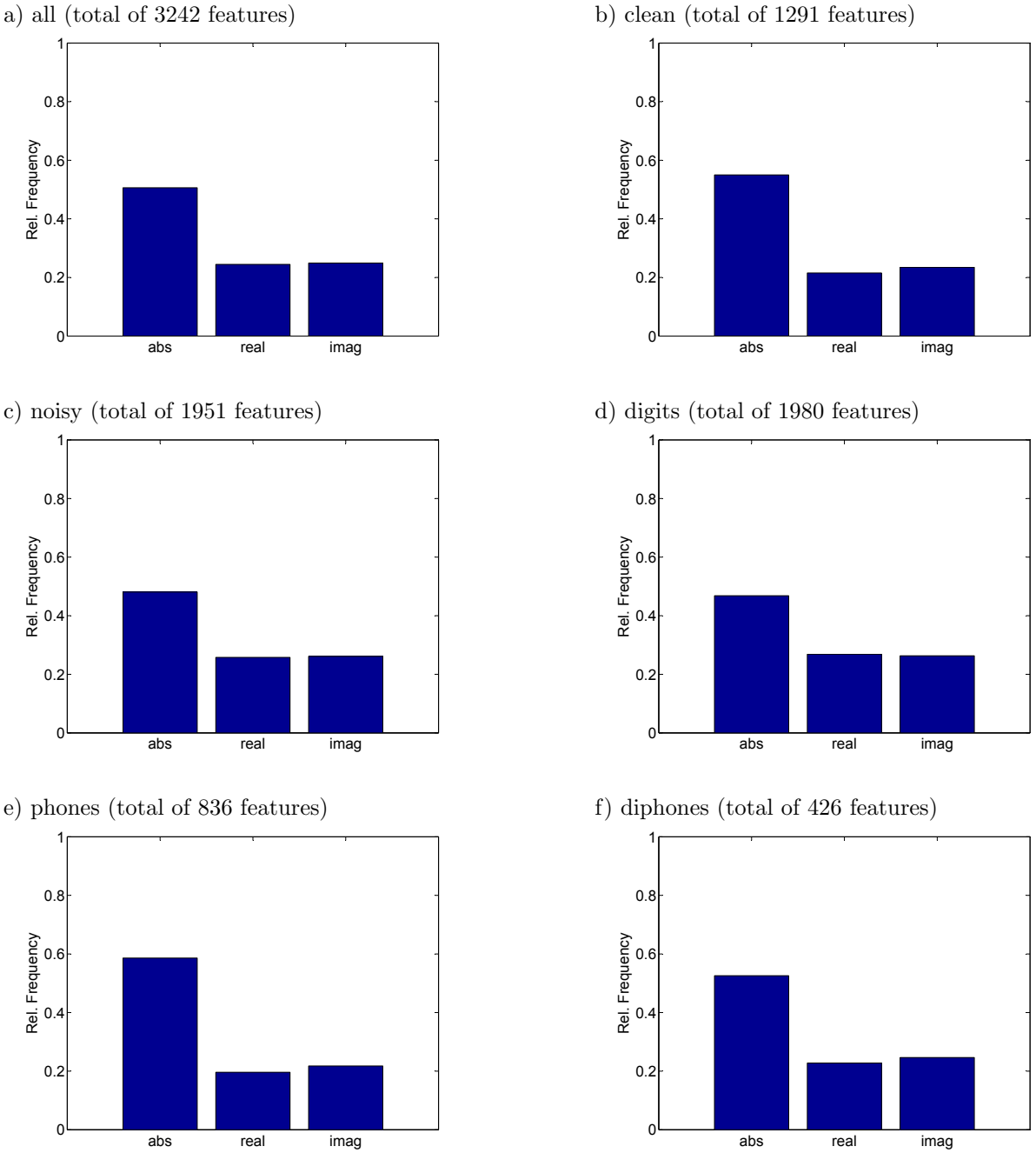
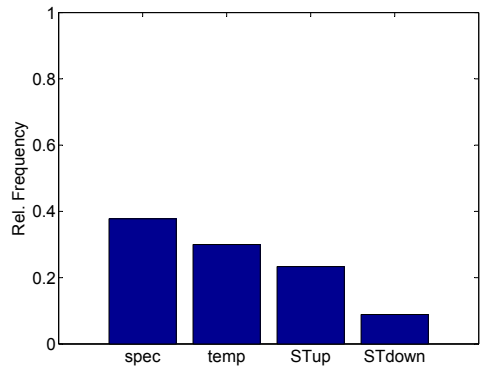
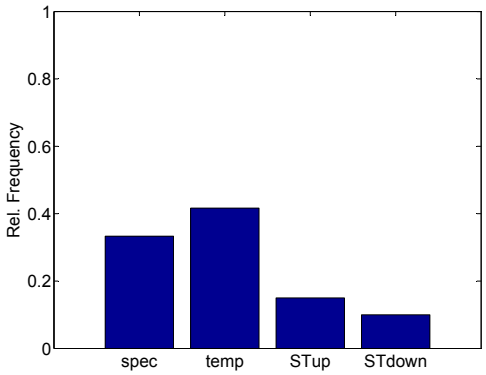


Figure A.1: Distribution of the three modes of Gabor features in the optimized sets. The relative frequency of complex ('abs'), real-valued cosine ('real', phase of $\pi/2$) and sine ('imag', zero phase) filter functions is shown for a) all optimization runs, b) optimization on clean data only, c) optimization on noisy data only as well as for d) digit targets only, e) phoneme targets only, and f) diphone targets only. See Chapter 7 for further description.

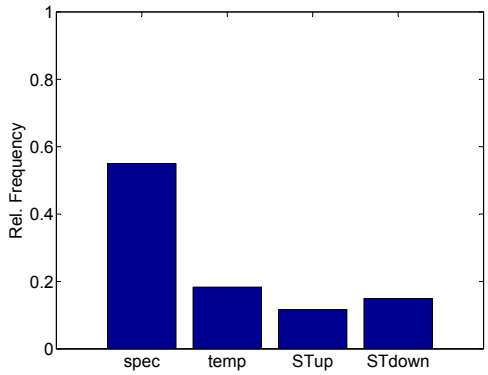
a) fricatives (total of 90 features)



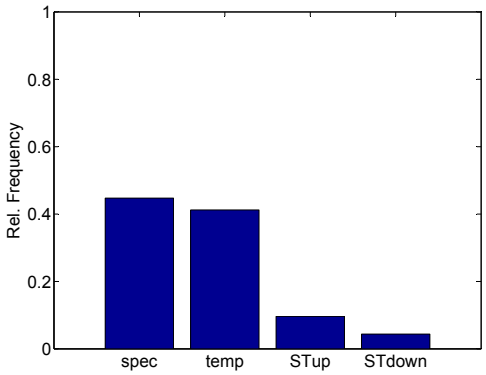
b) nasals (total of 60 features)



c) vowels (total of 120 features)



d) stops (total of 114 features)



e) diphthongs (total of 92 features)

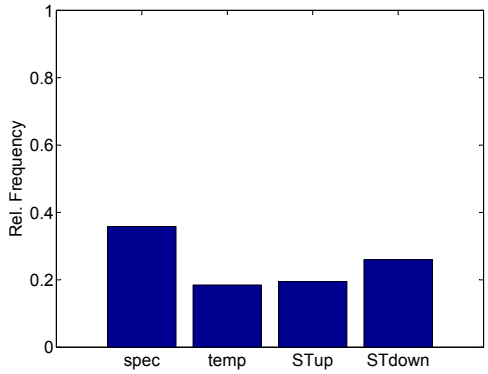


Figure A.2: Distribution of the four types of Gabor features in the optimized sets. The relative frequency of purely spectral ('spec'), purely temporal ('temp'), spectro-temporal ('ST') upwards ($\omega_t > 0$) and downwards ($\omega_t < 0$) directed Gabor modulation filters is shown for discrimination tasks within the group of a) fricatives, b) nasals, c) vowels, d) stops, and e) diphthongs. See Chapter 7 for further description.

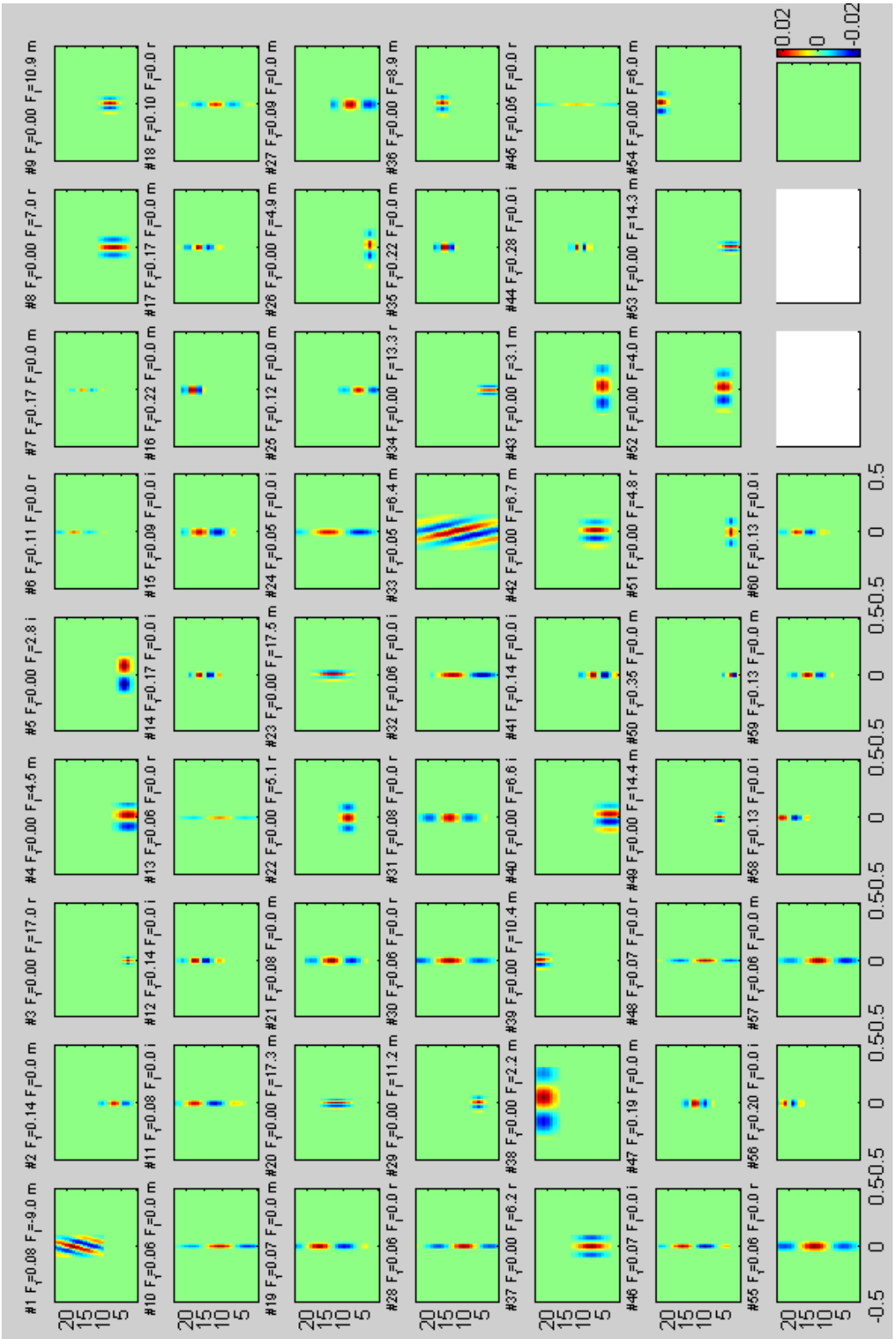


Figure A.3: Overview of Gabor set G1 (optimized on TIMIT phoneme inter-group discrimination). The 60 Gabor filter functions are plotted in a spectrogram of 23 channels times 1s. For the real valued filters ('r' for real and 'i' for imaginary part) color denotes the filter value at each point. For the complex case ('m'), the sum of real and imaginary components is plotted here. $\omega_f/2\pi$ in cycles/channel and $\omega_t/2\pi$ in Hz are in the title of each sub-plot. The full parameter set is given in Table A.1 on page 167. See Chapters 7 and 8 for further description.

Table A.1: Table of parameters for Gabor set G1 (optimized on TIMIT phoneme inter-group discrimination). # denotes the filter number as in Figure A.3, f_0 the number of the center frequency channel (on a scale from 1 to 23, low channel number equals low center frequency), ω_f and ω_t the spectral and temporal radian modulation frequency, respectively, σ_f and σ_t the widths of the Gaussian envelope, and Δ_f and Δ_t the extends of the support to both sides of the center. 'mode' specifies whether a filter is real with $\pi/2$ phase ('real'), zero phase ('imag') or complex ('mag'). 'type' highlights whether a filter is purely temporal, spectral or spectro-temporal ('ST'). See Chapter 7 for further description.

#	f_0	$\omega_f/2\pi$	$\omega_t/2\pi$	σ_f	σ_t	Δ_f	Δ_t	mode	type
		[cycl./chan.]	[100Hz]	[chan]	[10ms]	[chan]	[10ms]		
1	17	0.081	-0.090	6.141	5.556	7	10	mag	STup
2	6	0.140	0.000	3.571	1.540	5	2	mag	spectral
3	3	0.000	0.170	0.813	2.947	2	4	real	temporal
4	3	0.000	0.045	2.173	11.018	4	13	mag	temporal
5	4	0.000	0.028	1.246	17.758	2	19	imag	temporal
6	18	0.111	0.000	4.496	0.894	9	2	real	spectral
7	15	0.170	0.000	2.947	0.795	6	2	mag	spectral
8	7	0.000	0.070	2.795	7.114	4	10	real	temporal
9	8	0.000	0.109	1.546	4.587	4	9	mag	temporal
10	9	0.061	0.000	8.242	0.863	16	2	mag	spectral
11	15	0.084	0.000	5.965	1.598	11	2	imag	spectral
12	16	0.142	0.000	3.519	1.607	7	4	imag	spectral
13	11	0.064	0.000	7.760	0.822	16	2	real	spectral
14	15	0.172	0.000	2.912	1.863	4	3	imag	spectral
15	14	0.094	0.000	5.301	2.168	7	4	imag	spectral
16	18	0.224	0.000	2.230	2.435	3	5	mag	spectral
17	16	0.167	0.000	2.996	1.957	6	3	mag	spectral
18	12	0.102	0.000	4.905	1.416	10	3	real	spectral
19	15	0.069	0.000	7.298	1.786	11	4	mag	spectral
20	12	0.000	0.173	2.466	2.896	4	5	mag	temporal
21	12	0.085	0.000	5.890	2.267	8	4	mag	spectral
22	9	0.000	0.051	1.403	9.873	2	15	real	temporal
23	13	0.000	0.175	2.763	2.854	6	6	mag	temporal
24	10	0.053	0.000	9.394	1.612	18	3	imag	spectral
25	5	0.117	0.000	4.284	1.836	6	4	mag	spectral
26	3	0.000	0.049	0.696	10.287	2	19	mag	temporal
27	7	0.093	0.000	5.392	2.938	6	6	mag	spectral
28	10	0.064	0.000	7.823	1.503	13	2	real	spectral
29	6	0.000	0.112	1.084	4.464	2	9	mag	temporal
30	14	0.055	0.000	9.050	2.261	12	3	real	spectral
31	14	0.083	0.000	5.999	2.915	10	4	real	spectral
32	9	0.056	0.000	8.999	1.583	10	4	imag	spectral
33	11	0.050	0.064	10.087	7.756	12	15	mag	STdown
34	3	0.000	0.133	2.333	3.771	3	5	real	temporal
35	15	0.218	0.000	2.289	2.491	3	4	mag	spectral
36	16	0.000	0.089	0.956	5.640	2	12	mag	temporal
37	8	0.000	0.062	2.746	8.023	5	10	real	temporal
38	21	0.000	0.022	2.179	23.224	4	32	mag	temporal
39	22	0.000	0.104	1.627	4.796	3	8	mag	temporal
40	3	0.000	0.066	2.436	7.587	4	16	imag	temporal
41	6	0.143	0.000	3.487	2.485	5	3	imag	spectral
42	7	0.000	0.067	2.481	7.474	4	15	mag	temporal
43	5	0.000	0.031	1.246	16.274	3	22	mag	temporal
44	11	0.279	0.000	1.792	2.405	4	4	imag	spectral
45	12	0.046	0.000	10.897	0.689	12	2	real	spectral
46	13	0.075	0.000	6.676	0.968	9	2	imag	spectral
47	12	0.188	0.000	2.660	2.363	4	5	mag	spectral
48	10	0.068	0.000	7.350	0.867	14	2	real	spectral
49	6	0.000	0.144	0.930	3.466	1	6	mag	temporal
50	3	0.351	0.000	1.423	1.371	2	3	mag	spectral
51	3	0.000	0.048	0.725	10.441	1	13	real	temporal
52	5	0.000	0.040	1.519	12.485	2	19	mag	temporal
53	3	0.000	0.143	1.817	3.503	3	7	mag	temporal
54	22	0.000	0.060	1.145	8.328	2	12	mag	temporal
55	13	0.056	0.000	8.920	2.846	15	4	real	spectral
56	20	0.202	0.000	2.470	1.906	4	4	imag	spectral
57	10	0.060	0.000	8.287	2.343	17	3	mag	spectral
58	20	0.129	0.000	3.867	1.883	5	2	imag	spectral
59	14	0.128	0.000	3.913	1.939	6	3	mag	spectral
60	16	0.127	0.000	3.937	1.481	7	3	imag	spectral

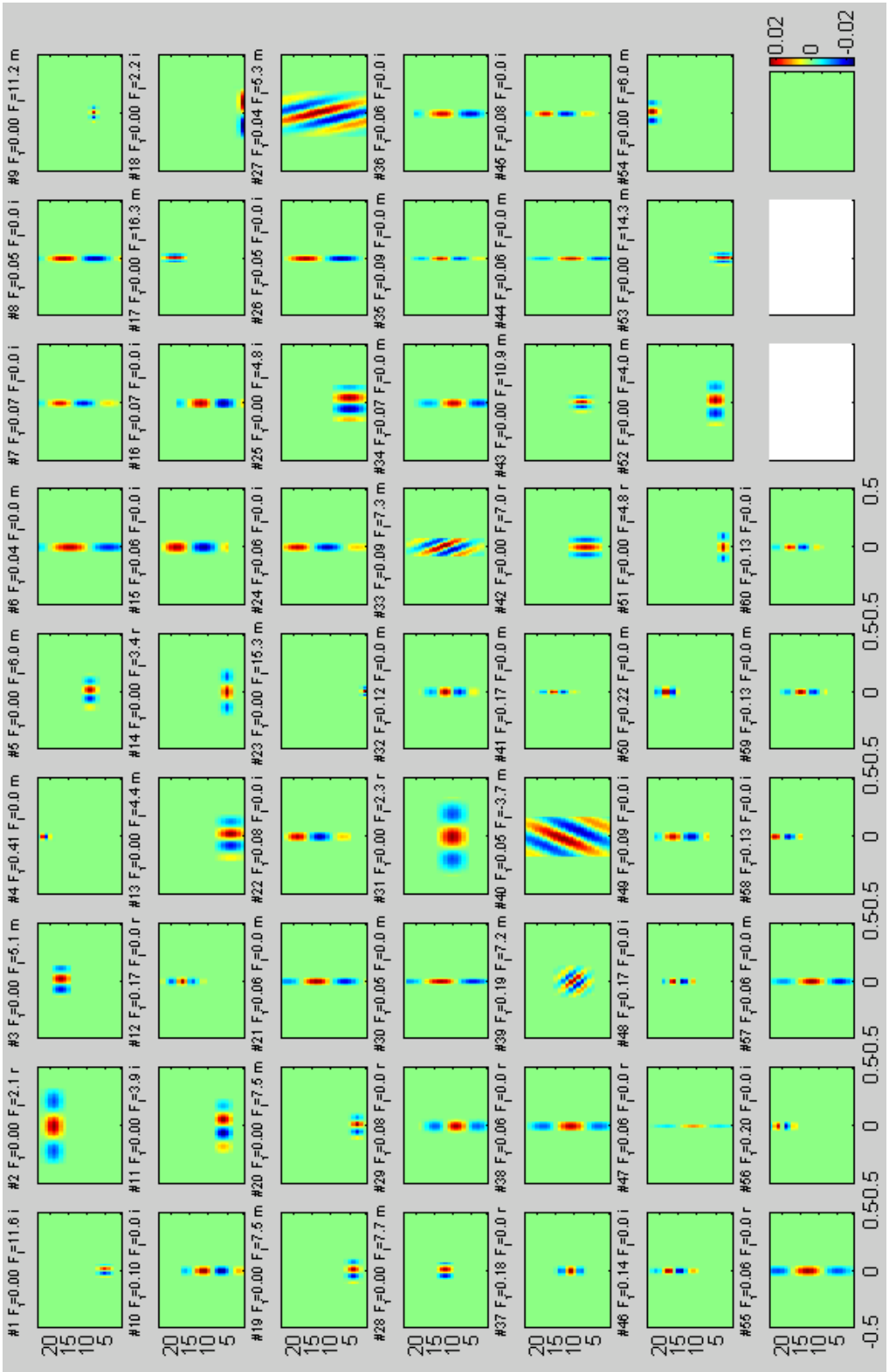


Figure A.4: Overview of Gabor set G2 (optimized on TIMIT phoneme inter-/within group discrimination). The 60 Gabor filter functions are plotted in a spectrogram of 23 channels times 1s. For the real valued filters ('r' for real and 'i' for imaginary part) color denotes the filter value at each point. For the complex case ('m'), the sum of real and imaginary components is plotted here. $\omega_f/2\pi$ in cycles/channel and $\omega_t/2\pi$ in Hz are in the title of each sub-plot. The full parameter set is given in Table A.2 on page 169. See Chapters 7 and 8 for further description.

Table A.2: Table of parameters for Gabor set G2 (optimized on TIMIT phoneme inter-/within group discrimination). # denotes the filter number as in Figure A.4, f_0 the number of the center frequency channel (on a scale from 1 to 23, low channel number equals low center frequency), ω_f and ω_t the spectral and temporal radian modulation frequency, respectively, σ_f and σ_t the widths of the Gaussian envelope, and Δ_f and Δ_t the extends of the support to both sides of the center. 'mode' specifies whether a filter is real with or $\pi/2$ phase ('real'), zero phase ('imag') or complex ('mag'). 'type' highlights whether a filter is purely temporal, spectral or spectro-temporal ('ST'). See Chapter 7 for further description.

#	f_0	$\omega_f/2\pi$	$\omega_t/2\pi$	σ_f	σ_t	Δ_f	Δ_t	mode	type
		[cycl./chan.]	[100Hz]	[chan]	[10ms]	[chan]	[10ms]		
1	5	0.000	0.116	1.182	4.309	2	5	imag	temporal
2	19	0.000	0.021	1.590	23.657	3	33	real	temporal
3	17	0.000	0.051	1.685	9.793	2	13	mag	temporal
4	22	0.406	0.000	1.233	1.401	3	3	mag	spectral
5	9	0.000	0.060	1.172	8.290	2	16	mag	temporal
6	12	0.044	0.000	11.392	2.418	21	4	mag	spectral
7	14	0.070	0.000	7.096	1.638	13	4	imag	spectral
8	12	0.053	0.000	9.429	1.800	11	4	imag	spectral
9	8	0.000	0.112	0.538	4.459	1	8	mag	temporal
10	9	0.095	0.000	5.257	2.589	8	4	imag	spectral
11	6	0.000	0.039	1.463	12.847	2	24	imag	temporal
12	17	0.166	0.000	3.013	2.005	6	4	real	spectral
13	4	0.000	0.044	2.280	11.301	4	21	mag	temporal
14	5	0.000	0.034	0.863	14.778	2	22	real	temporal
15	15	0.064	0.000	7.873	2.887	10	6	imag	spectral
16	9	0.074	0.000	6.716	2.972	9	4	imag	spectral
17	19	0.000	0.163	1.896	3.066	3	4	mag	temporal
18	1	0.000	0.022	0.575	22.854	2	24	imag	temporal
19	4	0.000	0.075	1.058	6.629	2	11	mag	temporal
20	3	0.000	0.075	0.977	6.684	2	12	mag	temporal
21	12	0.060	0.000	8.272	2.359	16	3	mag	spectral
22	16	0.076	0.000	6.600	2.327	11	3	imag	spectral
23	1	0.000	0.153	0.532	3.270	1	6	mag	temporal
24	15	0.057	0.000	8.703	1.998	16	3	imag	spectral
25	5	0.000	0.048	2.772	10.456	4	16	imag	temporal
26	12	0.047	0.000	10.561	1.760	15	3	imag	spectral
27	12	0.044	0.053	11.450	9.517	23	19	mag	STdown
28	12	0.000	0.077	1.345	6.453	2	7	mag	temporal
29	9	0.080	0.000	6.264	2.813	10	6	real	spectral
30	11	0.052	0.000	9.551	1.574	11	3	mag	spectral
31	10	0.000	0.023	2.272	21.552	4	34	real	temporal
32	11	0.120	0.000	4.170	2.725	8	4	mag	spectral
33	12	0.091	0.073	5.473	6.877	10	8	mag	STdown
34	8	0.068	0.000	7.355	2.241	13	4	mag	spectral
35	12	0.088	0.000	5.670	0.870	11	2	mag	spectral
36	9	0.060	0.000	8.362	1.934	11	3	imag	spectral
37	11	0.178	0.000	2.804	2.542	4	4	real	spectral
38	11	0.062	0.000	8.118	2.621	11	3	real	spectral
39	10	0.193	0.072	2.592	6.964	5	14	mag	STdown
40	12	0.053	-0.037	9.473	13.623	15	17	mag	STup
41	15	0.170	0.000	2.947	0.795	6	2	mag	spectral
42	7	0.000	0.070	2.795	7.114	4	10	real	temporal
43	8	0.000	0.109	1.546	4.587	4	9	mag	temporal
44	9	0.061	0.000	8.242	0.863	16	2	mag	spectral
45	15	0.084	0.000	5.965	1.598	11	2	imag	spectral
46	16	0.142	0.000	3.519	1.607	7	4	imag	spectral
47	11	0.064	0.000	7.760	0.822	16	2	real	spectral
48	15	0.172	0.000	2.912	1.863	4	3	imag	spectral
49	14	0.094	0.000	5.301	2.168	7	4	imag	spectral
50	18	0.224	0.000	2.230	2.435	3	5	mag	spectral
51	3	0.000	0.048	0.725	10.441	1	13	real	temporal
52	5	0.000	0.040	1.519	12.485	2	19	mag	temporal
53	3	0.000	0.143	1.817	3.503	3	7	mag	temporal
54	22	0.000	0.060	1.145	8.328	2	12	mag	temporal
55	13	0.056	0.000	8.920	2.846	15	4	real	spectral
56	20	0.202	0.000	2.470	1.906	4	4	imag	spectral
57	10	0.060	0.000	8.287	2.343	17	3	mag	spectral
58	20	0.129	0.000	3.867	1.883	5	2	imag	spectral
59	14	0.128	0.000	3.913	1.939	6	3	mag	spectral
60	16	0.127	0.000	3.937	1.481	7	3	imag	spectral

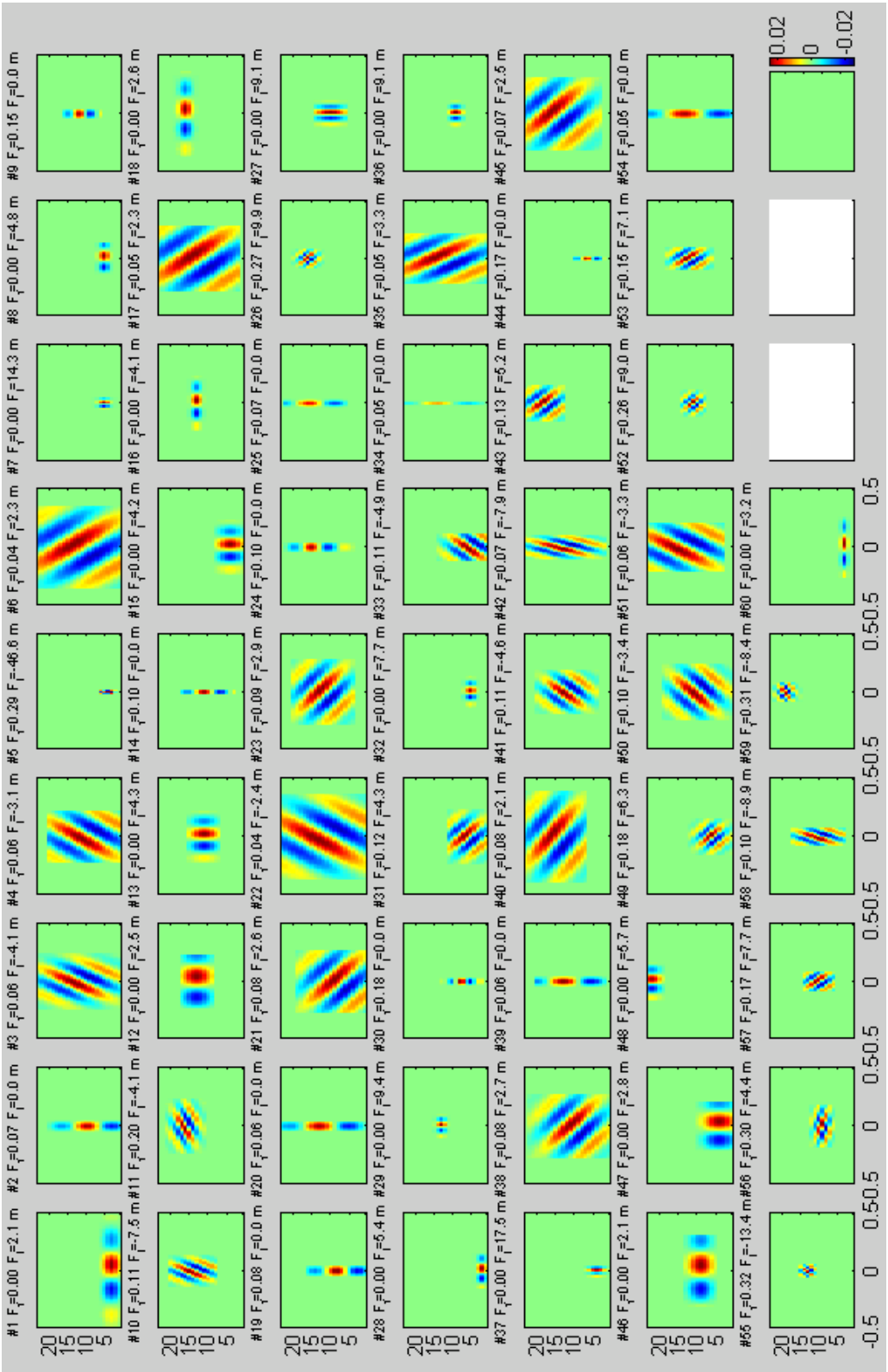


Figure A.5: Overview of Gabor set G3 (optimized on zifkom German digits). The 60 Gabor filter functions are plotted in a spectrogram of 23 channels times 1s. For the real valued filters ('r' for real and 'i' for imaginary part) color denotes the filter value at each point. For the complex case ('m'), the sum of real and imaginary components is plotted here. $\omega_f/2\pi$ in cycles/channel and $\omega_t/2\pi$ in Hz are in the title of each sub-plot. The full parameter set is given in Table A.3 on page 171. See Chapters 7 and 8 for further description.

Table A.3: Table of parameters for Gabor set G3 (optimized on zifkom German digits). # denotes the filter number as in Figure A.5, f_0 the number of the center frequency channel (on a scale from 1 to 23, low channel number equals low center frequency), ω_f and ω_t the spectral and temporal radian modulation frequency, respectively, σ_f and σ_t the widths of the Gaussian envelope, and Δ_f and Δ_t the extends of the support to both sides of the center. 'mode' specifies whether a filter is real with $\pi/2$ phase ('real'), zero phase ('imag') or complex ('mag'). 'type' highlights whether a filter is purely temporal, spectral or spectro-temporal ('ST'). See Chapter 7 for further description.

#	f_0	$\omega_f/2\pi$	$\omega_t/2\pi$	σ_f	σ_t	Δ_f	Δ_t	mode	type
		[cycl./chan.]	[100Hz]	[chan]	[10ms]	[chan]	[10ms]		
1	3	0.000	0.021	1.500	24.055	3	48	mag	temporal
2	8	0.068	0.000	7.387	2.300	12	5	mag	spectral
3	12	0.062	-0.041	8.076	12.089	12	24	mag	STup
4	10	0.057	-0.031	8.750	15.945	10	22	mag	STup
5	4	0.293	-0.466	1.705	1.073	2	2	mag	STup
6	12	0.044	0.023	11.401	21.589	13	36	mag	STdown
7	5	0.000	0.143	1.034	3.504	3	5	mag	temporal
8	5	0.000	0.048	0.964	10.466	2	13	mag	temporal
9	11	0.146	0.000	3.418	2.242	5	4	mag	spectral
10	14	0.110	-0.075	4.566	6.683	6	13	mag	STup
11	16	0.201	-0.041	2.488	12.265	5	24	mag	STup
12	13	0.000	0.025	2.643	20.152	4	23	mag	temporal
13	11	0.000	0.043	2.411	11.616	4	22	mag	temporal
14	10	0.104	0.000	4.799	1.620	7	2	mag	spectral
15	4	0.000	0.042	2.248	11.909	4	24	mag	temporal
16	13	0.000	0.041	0.821	12.207	2	24	mag	temporal
17	13	0.051	0.023	9.820	21.555	11	29	mag	STdown
18	16	0.000	0.026	1.201	18.952	3	37	mag	temporal
19	7	0.081	0.000	6.140	2.941	9	4	mag	spectral
20	11	0.058	0.000	8.610	2.550	12	4	mag	spectral
21	8	0.075	0.026	6.645	19.404	11	27	mag	STdown
22	12	0.044	-0.024	11.388	20.612	16	37	mag	STup
23	12	0.092	0.029	5.432	17.285	8	28	mag	STdown
24	14	0.103	0.000	4.859	2.446	10	3	mag	spectral
25	14	0.068	0.000	7.336	1.398	8	3	mag	spectral
26	16	0.267	0.099	1.875	5.047	4	8	mag	STdown
27	10	0.000	0.091	2.838	5.470	4	11	mag	temporal
28	2	0.000	0.054	0.748	9.252	2	15	mag	temporal
29	13	0.000	0.094	0.856	5.305	2	10	mag	temporal
30	7	0.183	0.000	2.737	2.073	6	3	mag	spectral
31	6	0.122	0.043	4.097	11.708	5	23	mag	STdown
32	5	0.000	0.077	0.968	6.498	2	13	mag	temporal
33	5	0.111	-0.049	4.497	10.123	9	12	mag	STup
34	12	0.054	0.000	9.195	0.756	16	2	mag	spectral
35	13	0.050	0.033	10.052	14.997	14	22	mag	STdown
36	9	0.000	0.091	1.165	5.515	2	11	mag	temporal
37	4	0.000	0.175	1.415	2.852	3	6	mag	temporal
38	10	0.082	0.027	6.100	18.262	12	28	mag	STdown
39	11	0.062	0.000	8.049	2.385	9	4	mag	spectral
40	15	0.081	0.021	6.168	23.383	8	39	mag	STdown
41	12	0.109	-0.046	4.592	10.957	8	19	mag	STup
42	13	0.073	-0.079	6.884	6.294	11	10	mag	STup
43	18	0.130	0.052	3.843	9.536	5	15	mag	STdown
44	6	0.166	0.000	3.006	0.773	4	1	mag	spectral
45	15	0.069	0.025	7.286	20.279	12	32	mag	STdown
46	9	0.000	0.021	2.114	23.523	4	31	mag	temporal
47	4	0.000	0.028	2.827	17.635	6	21	mag	temporal
48	22	0.000	0.057	1.590	8.825	3	16	mag	temporal
49	6	0.177	0.063	2.830	7.961	6	16	mag	STdown
50	10	0.095	-0.034	5.254	14.865	9	24	mag	STup
51	14	0.055	-0.033	9.055	14.940	11	21	mag	STup
52	11	0.261	0.090	1.918	5.584	3	11	mag	STdown
53	12	0.147	0.071	3.399	7.018	6	10	mag	STdown
54	11	0.049	0.000	10.184	2.385	21	3	mag	spectral
55	13	0.321	-0.134	1.556	3.744	2	5	mag	STup
56	9	0.303	0.044	1.652	11.268	3	18	mag	STdown
57	10	0.170	0.077	2.938	6.481	4	8	mag	STdown
58	10	0.100	-0.089	5.004	5.626	7	8	mag	STup
59	19	0.308	-0.084	1.623	5.975	4	8	mag	STup
60	3	0.000	0.032	0.554	15.556	1	26	mag	temporal

Table A.4: Baseline front end performance on Aurora 2 (TIDigits) and 3 (SpeechDat-car). As required for the ICSLP 2002 reference, the baseline system consisted of 13 mel-cepstral coefficients and additional delta and delta-delta features (ETSI ES 201 v1.1.2, 2000). The 39 features were used without quantization and the HTK back end had knowledge of the word boundaries ('endpointing'), producing perfect voice-activity detection.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.59	98.67	98.57	98.83	98.67	98.59	98.67	98.57	98.83	98.67	98.62	98.67	98.65	98.66
20 dB	97.82	97.94	98.24	97.47	97.87	97.73	97.61	97.61	97.66	97.65	97.67	97.58	97.63	97.73
15 dB	96.65	97.43	97.70	96.88	97.17	96.16	96.67	96.60	96.27	96.43	96.50	96.31	96.41	96.72
10 dB	94.38	95.47	96.18	94.11	95.04	92.94	94.86	93.71	93.92	93.86	93.83	93.92	93.88	94.33
5 dB	89.01	88.21	87.53	87.60	88.09	85.05	86.58	87.53	85.16	86.08	83.11	84.16	83.64	86.39
0 dB	67.85	63.18	54.10	63.71	62.21	60.88	63.06	66.27	58.07	62.07	46.21	56.35	51.28	59.97
-5dB	26.56	27.33	20.22	23.63	24.44	27.11	27.66	29.91	21.75	26.61	19.22	24.73	21.98	24.81
Average	89.14	88.45	86.75	87.95	88.07	86.55	87.76	88.34	86.22	87.22	83.46	85.66	84.56	87.03

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.89	99.03	99.05	99.26	99.06	98.89	99.03	99.05	99.26	99.06	99.17	99.09	99.13	99.07
20 dB	96.75	90.54	97.08	96.20	95.14	90.14	95.86	89.95	94.79	92.69	93.37	95.13	94.25	93.98
15 dB	91.53	72.19	88.55	90.03	85.58	74.52	88.15	73.84	81.24	79.44	86.03	89.09	87.56	83.52
10 dB	75.53	47.61	63.53	72.29	64.74	51.89	66.05	49.27	55.20	55.60	71.94	75.03	73.49	62.83
5 dB	47.34	22.91	30.75	39.08	35.02	26.80	36.28	24.60	24.96	28.16	50.63	50.57	50.60	35.39
0 dB	22.44	5.53	10.71	14.25	13.23	7.12	17.35	10.50	9.50	11.12	24.53	23.64	24.09	14.56
-5dB	10.65	0.12	6.83	6.85	6.11	0.95	8.62	5.28	6.14	5.25	12.90	11.19	12.05	6.95
Average	66.72	47.76	58.12	62.37	58.74	50.09	60.74	49.63	53.14	53.40	65.30	66.69	66.00	58.06

Aurora3 SpeechDatCar. Accuracy															
Finnish			Spanish			German			Danish			Average			
wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm
92.74	80.51	40.53	92.94	83.31	51.55	91.20	81.04	73.17	87.28	67.32	39.37	91.04	78.05	51.16

Table A.5: Absolute recognition results for the Tandem G1-R1-P system (Gabor set optimized on TIMIT phoneme inter-group discrimination, combined with mel-spectrogram based Tandem features via posterior combination) on Aurora 2 (TIDigits). See Chapter 7 for further description.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.23	99.00	98.93	99.11	99.07	99.23	99.00	98.93	99.11	99.07	99.20	99.03	99.12	99.08
20 dB	98.74	98.64	98.54	98.7	98.66	98.71	98.34	98.69	98.61	98.59	98.83	98.40	98.62	98.62
15 dB	97.88	98.00	98.18	97.93	98.00	97.88	97.64	97.79	97.9	97.80	97.85	97.46	97.66	97.85
10 dB	95.76	96.13	96.09	95.74	95.93	95.27	95.25	96.27	95.68	95.62	95.03	95.25	95.14	95.65
5 dB	89.68	91.05	90.61	88.46	89.95	87.29	87.45	89.5	88.65	88.22	87.38	86.06	86.72	88.61
0 dB	71.54	67.35	63.41	70.75	68.26	65.64	66.08	69.82	63.34	66.22	58.77	58.89	58.83	65.56
-5dB	31.32	24.52	19.00	31.53	26.59	27.85	26.30	28.06	20.98	25.80	21.74	23.43	22.59	25.47
Average	90.72	90.23	89.37	90.32	90.16	88.96	88.95	90.41	88.84	89.29	87.57	87.21	87.39	89.26

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.32	99.21	99.11	99.26	99.23	99.32	99.21	99.11	99.26	99.23	99.39	99.24	99.32	99.24
20 dB	98.31	97.67	98.36	97.99	98.08	97.27	97.88	98.03	98.03	97.80	98.00	97.85	97.93	97.94
15 dB	96.04	92.78	96.81	96.33	95.49	91.07	95.80	93.59	95.40	93.97	96.35	96.95	96.65	95.11
10 dB	88.95	77.63	87.24	91.08	86.23	75.68	88.85	80.58	86.70	82.95	90.79	91.41	91.10	85.89
5 dB	71.57	50.63	62.03	73.46	64.42	49.49	70.59	53.59	61.74	58.85	76.36	76.12	76.24	64.56
0 dB	40.1	19.89	24.84	39.56	31.10	20.51	38.36	24.96	26.10	27.48	45.59	46.31	45.95	32.62
-5dB	11.54	3.02	12.02	11.60	9.55	0.49	16.69	7.43	11.97	9.15	12.43	19.20	15.82	10.64
Average	78.99	67.72	73.86	79.68	75.06	66.80	78.30	70.15	73.59	72.21	81.42	81.73	81.57	75.22

Table A.6: Recognition results for the Tandem G1-R1-P system (Gabor set optimized on TIMIT phoneme inter-group discrimination, combined with mel-spectrogram based Tandem features via posterior combination) on Aurora 2 (TIDigits) relative to the baseline results (as in Table A.4). See Chapter 7 for further description.

Aurora2 TIDigits. Relative Improvement. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	45.39%	24.81%	25.17%	23.93%	29.83%	45.39%	24.81%	25.17%	23.93%	29.83%	42.03%	27.07%	34.55%	30.77%
20 dB	42.20%	33.98%	17.05%	48.62%	35.46%	43.17%	30.54%	45.19%	40.60%	39.88%	49.79%	33.88%	41.83%	38.50%
15 dB	36.72%	22.18%	20.87%	33.65%	28.35%	44.79%	29.13%	35.00%	43.70%	38.16%	38.57%	31.17%	34.87%	33.58%
10 dB	24.56%	14.57%	-2.36%	27.67%	16.11%	33.00%	7.59%	40.70%	28.95%	27.56%	19.45%	21.88%	20.66%	21.60%
5 dB	6.10%	24.09%	24.70%	6.94%	15.45%	14.98%	6.48%	15.80%	23.52%	15.20%	25.28%	11.99%	18.64%	15.99%
0 dB	11.48%	11.33%	20.28%	19.40%	15.62%	12.17%	8.18%	10.52%	12.57%	10.86%	23.35%	5.82%	14.58%	13.51%
-5dB	6.48%	-3.87%	-1.53%	10.34%	2.86%	1.02%	-1.88%	-2.64%	-0.98%	-1.12%	3.12%	-1.73%	0.70%	0.83%
Average	24.21%	21.23%	16.11%	27.26%	22.20%	29.62%	16.38%	29.44%	29.87%	26.33%	31.29%	20.95%	26.12%	24.64%

Aurora2 TIDigits. Relative Improvement. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	38.74%	18.56%	6.32%	0.00%	15.90%	38.74%	18.56%	6.32%	0.00%	15.90%	26.51%	16.48%	21.49%	17.02%
20 dB	48.00%	75.37%	43.84%	47.11%	53.58%	72.31%	48.79%	80.40%	62.19%	65.92%	69.83%	55.85%	62.84%	60.37%
15 dB	53.25%	74.04%	72.14%	63.19%	65.65%	64.95%	64.56%	75.50%	75.48%	70.12%	73.87%	72.04%	72.96%	68.90%
10 dB	54.84%	57.30%	65.01%	67.81%	61.24%	49.45%	67.16%	61.72%	70.31%	62.16%	67.18%	65.60%	66.39%	62.64%
5 dB	46.01%	35.96%	45.17%	56.43%	45.89%	31.00%	53.84%	38.45%	49.01%	43.08%	52.12%	51.69%	51.90%	45.97%
0 dB	22.77%	15.20%	15.82%	29.52%	20.83%	14.42%	25.42%	16.16%	18.34%	18.58%	27.91%	29.69%	28.80%	21.52%
-5dB	1.00%	2.90%	5.57%	5.10%	3.64%	-0.46%	8.83%	2.27%	6.21%	4.21%	-0.54%	9.02%	4.24%	3.99%
Average	44.97%	51.57%	48.40%	52.81%	49.44%	46.43%	51.95%	54.44%	55.07%	51.97%	58.18%	54.97%	56.58%	51.88%

Table A.7: Absolute recognition results for the Tandem G3-R1-P system (Gabor set optimized on zifkom German digits, combined with mel-spectrogram based Tandem features via posterior combination) on Aurora 2 (TIDigits). See Chapter 7 for further description.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.11	98.70	99.11	99.04	98.99	99.11	98.70	99.11	99.04	98.99	99.23	98.61	98.92	98.98
20 dB	99.08	98.55	98.60	98.67	98.73	98.53	98.10	98.60	98.73	98.49	98.99	98.13	98.56	98.60
15 dB	98.00	97.76	98.03	97.96	97.94	97.42	97.55	97.38	98.03	97.60	98.10	97.37	97.74	97.76
10 dB	96.01	95.50	96.12	95.22	95.71	94.32	94.65	95.53	95.40	94.98	95.79	95.16	95.48	95.37
5 dB	89.84	89.72	91.26	87.97	89.70	85.42	86.82	88.99	88.65	87.47	89.28	87.15	88.22	88.51
0 dB	72.40	65.81	66.72	69.92	68.71	62.30	65.72	70.18	65.44	65.91	65.24	62.82	64.03	66.66
-5dB	34.02	24.49	22.01	32.21	28.18	25.94	27.39	28.60	22.09	26.01	25.91	25.79	25.85	26.85
Average	91.07	89.47	90.15	89.95	90.16	87.60	88.57	90.14	89.25	88.89	89.48	88.13	88.80	89.38

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.42	99.09	99.25	99.20	99.24	99.42	99.09	99.25	99.20	99.24	99.48	99.09	99.29	99.25
20 dB	98.86	96.34	98.75	98.52	98.12	96.16	97.67	97.76	98.06	97.41	98.99	97.91	98.45	97.90
15 dB	97.61	90.05	97.08	96.82	95.39	88.98	95.62	93.26	95.28	93.29	97.54	96.52	97.03	94.88
10 dB	92.45	72.34	88.43	91.21	86.11	72.18	87.03	79.81	85.87	81.22	93.43	91.66	92.55	85.44
5 dB	76.14	46.10	61.94	72.72	64.23	47.80	67.84	52.19	59.77	56.90	82.99	77.00	80.00	64.45
0 dB	45.69	17.02	28.21	36.32	31.81	19.68	37.73	23.95	28.14	27.38	54.13	45.86	50.00	33.67
-5dB	17.90	1.84	10.86	10.34	10.24	0.52	17.17	6.71	11.08	8.87	20.76	19.74	20.25	11.69
Average	82.15	64.37	74.88	79.12	75.13	64.96	77.18	69.39	73.42	71.24	85.42	81.79	83.60	75.27

Table A.8: Recognition results for the Tandem G3-R1-P system (Gabor set optimized on zifkom German digits, combined with mel-spectrogram based Tandem features via posterior combination) on Aurora 2 (TIDigits) relative to the baseline results (as in Table A.4). See Chapter 7 for further description

Aurora2 TIDigits. Relative Improvement. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	36.88%	2.26%	37.76%	17.95%	23.71%	36.88%	2.26%	37.76%	17.95%	23.71%	44.20%	-4.51%	19.85%	22.94%
20 dB	57.80%	29.61%	20.45%	47.43%	38.82%	35.24%	20.50%	41.42%	45.73%	35.72%	56.65%	22.73%	39.69%	37.76%
15 dB	40.30%	12.84%	14.35%	34.62%	25.53%	32.81%	26.43%	22.94%	47.18%	32.34%	45.71%	28.73%	37.22%	30.59%
10 dB	29.00%	0.66%	-1.57%	18.85%	11.74%	19.55%	-4.09%	28.93%	24.34%	17.18%	31.77%	20.39%	26.08%	16.78%
5 dB	7.55%	12.81%	29.91%	2.98%	13.31%	2.47%	1.79%	11.71%	23.52%	9.87%	36.53%	18.88%	27.70%	14.82%
0 dB	14.15%	7.14%	27.49%	17.11%	16.48%	3.63%	7.20%	11.59%	17.58%	10.00%	35.38%	14.82%	25.10%	15.61%
-5dB	10.16%	-3.91%	2.24%	11.23%	4.93%	-1.61%	-0.37%	-1.87%	0.43%	-0.85%	8.28%	1.41%	4.85%	2.60%
Average	29.76%	12.61%	18.13%	24.20%	21.17%	18.74%	10.37%	23.32%	31.67%	21.02%	41.21%	21.11%	31.16%	23.11%

Aurora2 TIDigits. Relative Improvement. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	47.75%	6.19%	21.05%	-8.11%	16.72%	47.75%	6.19%	21.05%	-8.11%	16.72%	37.35%	0.00%	18.67%	17.11%
20 dB	64.92%	61.31%	57.19%	61.05%	61.12%	61.05%	43.72%	77.71%	62.76%	61.31%	84.77%	57.08%	70.93%	63.16%
15 dB	71.78%	64.22%	74.50%	68.10%	69.65%	56.75%	63.04%	74.24%	74.84%	67.22%	82.39%	68.10%	75.25%	69.80%
10 dB	69.15%	47.20%	68.28%	68.28%	63.23%	42.17%	61.80%	60.20%	68.46%	58.16%	76.59%	66.60%	71.59%	62.87%
5 dB	54.69%	30.08%	45.04%	55.22%	46.26%	28.69%	49.53%	36.59%	46.39%	40.30%	65.55%	53.47%	59.51%	46.52%
0 dB	29.98%	12.16%	19.60%	25.74%	21.87%	13.52%	24.66%	15.03%	20.60%	18.45%	39.22%	29.10%	34.16%	22.96%
-5dB	8.11%	1.72%	4.33%	3.75%	4.48%	-0.43%	9.36%	1.51%	5.26%	3.92%	9.02%	9.63%	9.33%	5.23%
Average	58.10%	43.00%	52.92%	55.68%	52.42%	40.44%	48.55%	52.75%	54.61%	49.09%	69.70%	54.87%	62.29%	53.06%

Table A.9: Aurora 2 (TIDigits): Summary of recognition performance of the Tandem G1-R1-P system (Gabor set optimized on TIMIT phoneme inter-group discrimination, combined with mel-spectrogram based Tandem features via posterior combination) absolute and relative to the baseline results (as in Table A.4). See Chapter 7 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	9.84%	10.71%	12.61%	10.74%
Clean	24.94%	27.79%	18.43%	24.78%
Average	17.39%	19.25%	15.52%	17.76%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	22.20%	26.33%	26.12%	24.64%
Clean	49.44%	51.97%	56.58%	51.88%
Average	35.82%	39.15%	41.35%	38.26%

Table A.10: Aurora 2 (TIDigits): Summary of recognition performance of the Tandem G3-R1-P system (Gabor set optimized on zifkom German digits, combined with mel-spectrogram based Tandem features via posterior combination) absolute and relative to the baseline results (as in Table A.4). See Chapter 7 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	9.84%	10.71%	12.61%	10.74%
Clean	24.94%	27.79%	18.43%	24.78%
Average	17.39%	19.25%	15.52%	17.76%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	22.20%	26.33%	26.12%	24.64%
Clean	49.44%	51.97%	56.58%	51.88%
Average	35.82%	39.15%	41.35%	38.26%

Table A.11: Absolute recognition results for the Tandem G1-D system (Gabor set optimized on TIMIT phoneme inter-group discrimination, concatenated to Tandem Gabor features optimized on diphone targets) on Aurora 2 (TIDigits). See Chapter 7 for further description.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.02	99.15	98.90	99.32	99.10	99.02	99.15	98.90	99.32	99.10	99.17	99.00	99.09	99.10
20 dB	98.65	98.64	98.63	98.73	98.66	98.59	98.37	98.69	98.49	98.54	98.74	98.37	98.56	98.59
15 dB	97.88	97.88	98.03	97.75	97.89	96.65	97.46	97.44	97.59	97.29	97.42	97.31	97.37	97.54
10 dB	95.12	95.62	96.09	95.37	95.55	93.86	95.16	95.71	94.88	94.90	95.15	95.25	95.20	95.22
5 dB	89.28	89.39	90.96	88.43	89.52	84.16	88.09	88.43	87.13	86.95	87.14	86.49	86.82	87.95
0 dB	68.62	65.18	67.52	70.81	68.03	59.81	66.14	68.48	63.68	64.53	60.70	57.35	59.03	64.83
-5dB	30.12	23.25	21.83	31.93	26.78	22.29	29.32	29.38	20.70	25.42	21.71	23.64	22.68	25.42
Average	89.91	89.34	90.25	90.22	89.93	86.61	89.04	89.75	88.35	88.44	87.83	86.95	87.39	88.83

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.17	99.09	99.11	99.29	99.17	99.17	99.09	99.11	99.29	99.17	99.26	99.18	99.22	99.18
20 dB	98.13	97.97	98.36	98.18	98.16	95.92	97.88	97.67	97.28	97.19	98.19	97.61	97.90	97.72
15 dB	96.07	92.90	97.29	95.99	95.56	87.60	96.16	93.71	94.97	93.11	96.25	95.98	96.12	94.69
10 dB	89.25	78.60	91.89	90.77	87.63	71.81	89.54	81.60	89.23	83.05	91.13	90.90	91.02	86.47
5 dB	73.35	49.03	67.91	73.80	66.02	43.26	70.07	52.58	64.83	57.69	75.35	74.27	74.81	64.45
0 dB	41.05	15.45	24.69	36.22	29.35	13.33	37.12	21.80	26.26	24.63	43.81	42.35	43.08	30.21
-5dB	13.75	-2.39	8.98	9.60	7.49	-4.97	15.48	4.98	8.92		11.30	16.48	13.89	
Average	79.57	66.79	76.03	78.99	75.35	62.38	78.15	69.47	74.51	71.13	80.95	80.22	80.58	74.71

Table A.12: Recognition results for the Tandem G1-D system (Gabor set optimized on TIMIT phoneme inter-group discrimination, concatenated to Tandem Gabor features optimized on diphone targets) on Aurora 2 (TIDigits) relative to the baseline results (as in Table A.4). See Chapter 7 for further description.

Aurora2 TIDigits. Relative Improvement. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	30.50%	36.09%	23.08%	41.88%	32.89%	30.50%	36.09%	23.08%	41.88%	32.89%	39.86%	24.81%	32.33%	32.78%
20 dB	38.07%	33.98%	22.16%	49.80%	36.00%	37.89%	31.80%	45.19%	35.47%	37.59%	45.92%	32.64%	39.28%	37.29%
15 dB	36.72%	17.51%	14.35%	27.88%	24.11%	12.76%	23.72%	24.71%	35.39%	24.14%	26.29%	27.10%	26.69%	24.64%
10 dB	13.17%	3.31%	-2.36%	21.39%	8.88%	13.03%	5.84%	31.80%	15.79%	16.61%	21.39%	21.88%	21.63%	14.52%
5 dB	2.46%	10.01%	27.51%	6.69%	11.67%	-5.95%	11.25%	7.22%	13.27%	6.45%	23.86%	14.71%	19.28%	11.10%
0 dB	2.40%	5.43%	29.24%	19.56%	14.16%	-2.74%	8.34%	6.55%	13.38%	6.38%	26.94%	2.29%	14.61%	11.14%
-5dB	4.85%	-5.61%	2.02%	10.87%	3.03%	-6.61%	2.29%	-0.76%	-1.34%	-1.60%	3.08%	-1.45%	0.82%	0.73%
Average	18.56%	14.05%	18.18%	25.07%	18.96%	11.00%	16.19%	23.09%	22.66%	18.24%	28.88%	19.72%	24.30%	19.74%

Aurora2 TIDigits. Relative Improvement. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	25.23%	6.19%	6.32%	4.05%	10.45%	25.23%	6.19%	6.32%	4.05%	10.45%	10.84%	9.89%	10.37%	10.43%
20 dB	42.46%	78.54%	43.84%	52.11%	54.24%	58.62%	48.79%	76.82%	47.79%	58.01%	72.70%	50.92%	61.81%	57.26%
15 dB	53.60%	74.47%	76.33%	59.78%	66.05%	51.33%	67.59%	75.96%	73.19%	67.02%	73.16%	63.15%	68.15%	66.86%
10 dB	56.07%	59.15%	77.76%	66.69%	64.92%	41.41%	69.19%	63.73%	75.96%	62.57%	68.39%	63.56%	65.97%	64.19%
5 dB	49.39%	33.88%	53.66%	56.99%	48.48%	22.49%	53.03%	37.11%	53.13%	41.44%	50.07%	47.95%	49.01%	45.77%
0 dB	23.99%	10.50%	15.66%	25.62%	18.94%	6.69%	23.92%	12.63%	18.52%	15.44%	25.55%	24.50%	25.02%	18.76%
-5dB	3.47%	-2.51%	2.31%	2.95%	1.55%	-5.98%	7.51%	-0.32%	2.96%	1.04%	-1.84%	5.96%	2.06%	1.45%
Average	45.10%	51.31%	53.45%	52.24%	50.53%	36.11%	52.51%	53.25%	53.72%	48.89%	57.97%	50.02%	53.99%	50.57%

Table A.13: Absolute recognition results for the Tandem G1-R0-Q system (Gabor set optimized on TIMIT phoneme inter-group discrimination, concatenated to Aurora baseline features) on Aurora 2 (TIDigits). See Chapter 7 for further description.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.23	99.33	99.16	99.48	99.30	99.23	99.33	99.16	99.48	99.30	99.39	99.21	99.30	99.30
20 dB	98.99	98.85	98.75	98.92	98.88	98.80	98.55	98.93	99.29	98.89	98.77	98.58	98.68	98.84
15 dB	98.34	98.52	98.51	98.09	98.37	97.82	97.97	98.09	98.06	97.99	97.79	97.82	97.81	98.10
10 dB	96.56	96.10	96.78	95.80	96.31	94.81	95.83	96.12	95.62	95.60	95.61	95.92	95.77	95.92
5 dB	90.97	89.63	92.84	90.13	90.89	85.32	89.42	88.85	86.76	87.59	88.52	89.42	88.97	89.19
0 dB	72.92	62.88	70.50	73.99	70.07	56.09	69.53	68.71	63.68	64.50	60.15	63.42	61.79	66.19
-5dB	30.95	20.80	23.20	32.61	26.89	14.83	29.38	24.84	20.73	22.45	21.03	25.45	23.24	24.38
Average	91.56	89.20	91.48	91.39	90.90	86.57	90.26	90.14	88.68	88.91	88.17	89.03	88.60	89.65

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.45	99.40	99.37	99.54	99.44	99.45	99.40	99.37	99.54	99.44	99.48	99.37	99.43	99.44
20 dB	98.46	98.25	98.81	98.15	98.42	98.13	98.10	98.45	98.49	98.29	97.61	97.79	97.70	98.22
15 dB	96.22	94.20	96.93	96.70	96.01	93.00	96.40	95.17	95.28	94.96	94.81	96.34	95.58	95.51
10 dB	87.10	79.41	86.01	89.14	85.42	77.00	87.42	82.08	83.22	82.43	88.33	91.69	90.01	85.14
5 dB	64.45	45.13	54.16	66.12	57.47	45.29	60.61	49.90	51.34	51.79	69.27	73.67	71.47	57.99
0 dB	30.73	12.64	18.04	23.88	21.32	12.96	26.93	20.46	18.45	19.70	34.73	38.15	36.44	23.70
-5dB	12.19	-1.90	6.86	6.26	5.85	-5.10	10.55	3.52	5.65		11.54	14.90	13.22	
Average	75.39	65.93	70.79	74.80	71.73	65.28	73.89	69.21	69.36	69.43	76.95	79.53	78.24	72.11

Table A.14: Recognition results for the Tandem G1-R0-Q system (Gabor set optimized on TIMIT phoneme inter-group discrimination, concatenated to Aurora baseline features) on Aurora 2 (TIDigits) relative to the baseline results (as in Table A.4). See Chapter 7 for further description

Aurora2 TIDigits. Relative Improvement. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	45.39%	49.62%	41.26%	55.56%	47.96%	45.39%	49.62%	41.26%	55.56%	47.96%	55.80%	40.60%	48.20%	48.01%
20 dB	53.67%	44.17%	28.98%	57.31%	46.03%	47.14%	39.33%	55.23%	69.66%	52.84%	47.21%	41.32%	44.27%	48.40%
15 dB	50.45%	42.41%	35.22%	38.78%	41.71%	43.23%	39.04%	43.82%	47.99%	43.52%	36.86%	40.92%	38.89%	41.87%
10 dB	38.79%	13.91%	15.71%	28.69%	24.27%	26.49%	18.87%	38.31%	27.96%	27.91%	28.85%	32.89%	30.87%	27.05%
5 dB	17.83%	12.04%	42.58%	20.40%	23.22%	1.81%	21.16%	10.59%	10.78%	11.08%	32.03%	33.21%	32.62%	20.24%
0 dB	15.77%	-0.81%	35.73%	28.33%	19.75%	-12.24%	17.51%	7.23%	13.38%	6.47%	25.92%	16.20%	21.06%	14.70%
-5dB	5.98%	-8.99%	3.74%	11.76%	3.12%	-16.85%	2.38%	-7.23%	-1.30%	-5.75%	2.24%	0.96%	1.60%	-0.73%
Average	35.30%	22.34%	31.64%	34.70%	31.00%	21.28%	27.18%	31.04%	33.95%	28.36%	34.17%	32.91%	33.54%	30.45%

Aurora2 TIDigits. Relative Improvement. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	50.45%	38.14%	33.68%	37.84%	40.03%	50.45%	38.14%	33.68%	37.84%	40.03%	37.35%	30.77%	34.06%	38.84%
20 dB	52.62%	81.50%	59.25%	51.32%	61.17%	81.03%	54.11%	84.58%	71.02%	72.68%	63.95%	54.62%	59.29%	65.40%
15 dB	55.37%	79.14%	73.19%	66.90%	68.65%	72.53%	69.62%	81.54%	74.84%	74.63%	62.85%	66.45%	64.65%	70.24%
10 dB	47.28%	60.70%	61.64%	60.81%	57.61%	52.19%	62.95%	64.68%	62.54%	60.59%	58.41%	66.72%	62.57%	59.79%
5 dB	32.49%	28.82%	33.81%	44.39%	34.88%	25.26%	38.18%	33.55%	35.15%	33.04%	37.76%	46.73%	42.24%	35.61%
0 dB	10.69%	7.53%	8.21%	11.23%	9.41%	6.29%	11.59%	11.13%	9.89%	9.72%	13.52%	19.00%	16.26%	10.91%
-5dB	1.72%	-2.02%	0.03%	-0.63%	-0.23%	-6.11%	2.11%	-1.86%	-0.52%	-1.59%	-1.56%	4.18%	1.31%	-0.47%
Average	39.69%	51.54%	47.22%	46.93%	46.34%	47.46%	47.29%	55.09%	50.69%	50.13%	47.30%	50.71%	49.00%	48.39%

Table A.15: Aurora 2 (TIDigits): Summary of recognition performance of the Tandem G1-D system (Gabor set optimized on TIMIT phoneme inter-group discrimination, concatenated to Tandem Gabor features optimized on diphone targets) absolute and relative to the baseline results (as in Table A.4). See Chapter 7 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	10.07%	11.56%	12.61%	11.17%
Clean	24.66%	28.87%	19.42%	25.29%
Average	17.36%	20.21%	16.01%	18.23%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	18.96%	18.24%	24.30%	19.74%
Clean	50.53%	48.89%	53.99%	50.57%
Average	34.74%	33.56%	39.15%	35.15%

Table A.16: Aurora 2 (TIDigits): Summary of recognition performance of the Tandem G1-R0-Q system (Gabor set optimized on TIMIT phoneme inter-group discrimination, concatenated to Aurora baseline features) absolute and relative to the baseline results (as in Table A.4). See Chapter 7 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	9.10%	11.09%	11.40%	10.35%
Clean	28.27%	30.57%	21.76%	27.89%
Average	18.69%	20.83%	16.58%	19.12%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	31.00%	28.36%	33.54%	30.45%
Clean	46.34%	50.13%	49.00%	48.39%
Average	38.67%	39.25%	41.27%	39.42%

Table A.17: Absolute recognition results for the Tandem G1d system (Gabor set optimized on TIMIT phoneme inter-group discrimination and combined with the Qualcomm-ICSI-OGI feature stream) on Aurora 2 (TIDigits) and 3 (SpeechDat-car). See Chapter 8 for further description.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.39	99.24	99.16	99.48	99.32	99.39	99.24	99.16	99.48	99.32	99.36	99.46	99.41	99.34
20 dB	98.96	99.00	99.14	99.11	99.05	98.50	99.03	98.84	98.95	98.83	98.80	98.52	98.66	98.89
15 dB	98.46	98.37	98.33	98.58	98.44	98.04	98.28	98.42	98.43	98.29	98.59	97.67	98.13	98.32
10 dB	96.68	96.58	97.02	96.42	96.68	95.52	95.98	96.48	96.48	96.12	95.98	95.44	95.71	96.26
5 dB	91.89	91.48	93.02	89.66	91.51	88.55	89.54	90.75	90.19	89.76	90.36	88.63	89.50	90.41
0 dB	75.84	69.62	76.89	73.00	73.84	64.94	72.70	75.04	72.76	71.36	71.57	66.43	69.00	71.88
-5dB	42.03	29.32	37.46	40.45	37.32	29.14	37.33	36.12	37.10	34.92	34.76	32.34	33.55	35.61
Average	92.37	91.01	92.88	91.35	91.90	89.11	91.11	91.91	91.36	90.87	91.06	89.34	90.20	91.15

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.60	99.55	99.34	99.63	99.53	99.60	99.55	99.34	99.63	99.53	99.60	99.49	99.55	99.53
20 dB	99.17	98.88	99.19	99.07	99.08	98.43	98.85	98.96	98.89	98.78	99.11	98.61	98.86	98.92
15 dB	98.07	97.46	98.30	97.75	97.90	96.25	97.85	97.70	97.72	97.38	97.97	97.16	97.57	97.62
10 dB	95.03	93.14	96.36	95.03	94.89	90.70	94.14	94.21	94.88	93.48	94.60	94.26	94.43	94.24
5 dB	87.87	80.96	90.13	86.30	86.32	76.14	85.01	83.87	85.37	82.60	87.23	83.01	85.12	84.59
0 dB	66.26	50.21	67.43	63.44	61.84	45.53	61.85	58.84	63.38	57.40	59.87	57.51	58.69	59.43
-5dB	29.54	15.99	25.89	28.57	25.00	13.45	27.28	23.68	27.19	22.90	25.12	24.86	24.99	24.16
Average	89.28	84.13	90.28	88.32	88.00	81.41	87.54	86.72	88.05	85.93	87.76	86.11	86.93	86.96

Aurora3 SpeechDat-Car. Accuracy															
Finnish			Spanish			German			Danish			Average			
wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm
97.53	91.18	86.40	97.90	95.05	86.47	94.77	88.29	88.21	93.59	80.99	79.93	95.95	88.88	85.25

Table A.18: Recognition performance of the Tandem G1d system (Gabor set optimized on TIMIT phoneme inter-group discrimination and combined with the Qualcomm-ICSI-OGI feature stream) on Aurora 2 (TIDigits) and 3 (SpeechDat-car) relative to the baseline results (as in Table A.4). See Chapter 8 for further description.

Aurora2 TIDigits. Relative Improvement. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	56.74%	42.86%	41.26%	55.56%	49.10%	56.74%	42.86%	41.26%	55.56%	49.10%	53.62%	59.40%	56.51%	50.58%
20 dB	52.29%	51.46%	51.14%	64.82%	54.93%	33.92%	59.41%	51.46%	55.13%	49.98%	48.50%	38.84%	43.67%	50.70%
15 dB	54.03%	36.58%	27.39%	54.49%	43.12%	48.96%	48.35%	53.53%	57.91%	52.19%	59.71%	36.86%	48.29%	47.78%
10 dB	40.93%	24.50%	21.99%	39.22%	31.66%	36.54%	21.79%	44.04%	42.11%	36.12%	34.85%	25.00%	29.92%	33.10%
5 dB	26.21%	27.74%	44.03%	16.61%	28.64%	23.41%	22.06%	25.82%	33.89%	26.30%	42.92%	28.22%	35.57%	29.09%
0 dB	24.85%	17.49%	49.65%	25.60%	29.40%	10.38%	26.10%	26.00%	35.03%	24.38%	47.15%	23.09%	35.12%	28.53%
-5dB	21.06%	2.74%	21.61%	22.02%	16.86%	2.79%	13.37%	8.86%	19.62%	11.16%	19.24%	10.11%	14.67%	14.14%
Average	39.66%	31.55%	38.84%	40.15%	37.55%	30.64%	35.54%	40.17%	44.81%	37.79%	46.63%	30.40%	38.51%	37.84%

Aurora2 TIDigits. Relative Improvement. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	63.96%	53.61%	30.53%	50.00%	49.52%	63.96%	53.61%	30.53%	50.00%	49.52%	51.81%	43.96%	47.88%	49.20%
20 dB	74.46%	88.16%	72.26%	75.53%	77.60%	84.08%	72.22%	89.65%	78.69%	81.16%	86.58%	71.46%	79.02%	79.31%
15 dB	77.21%	90.87%	85.15%	77.43%	82.67%	85.28%	81.86%	91.21%	87.85%	86.55%	85.47%	73.97%	79.72%	83.63%
10 dB	79.69%	86.91%	90.02%	82.06%	84.67%	80.67%	82.74%	88.59%	88.57%	85.14%	80.76%	77.01%	78.88%	83.70%
5 dB	76.97%	75.30%	85.75%	77.51%	78.88%	67.40%	76.48%	78.61%	80.50%	75.75%	74.13%	65.63%	69.88%	75.83%
0 dB	56.50%	47.30%	63.52%	57.36%	56.17%	41.35%	53.84%	54.01%	59.54%	52.19%	46.83%	44.36%	45.59%	52.46%
-5dB	21.14%	15.89%	20.46%	23.32%	20.20%	12.62%	20.42%	19.43%	22.43%	18.72%	14.03%	15.39%	14.71%	18.51%
Average	72.97%	77.71%	79.34%	73.98%	76.00%	71.76%	73.43%	80.41%	79.03%	76.16%	74.75%	66.48%	70.62%	74.99%

Aurora3 SpeechDat-Car. Relative Improvement.															
Finnish			Spanish			German			Danish			Average		
wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm
65.98%	54.75%	77.13%	70.25%	70.34%	72.07%	40.57%	38.24%	56.06%	49.61%	41.83%	66.90%	56.60%	51.29%	68.04%

Table A.19: Aurora 2 (TIDigits): Summary of recognition performance of the Tandem G1d system (Gabor set optimized on TIMIT phoneme inter-group discrimination and combined with the Qualcomm-ICSI-OGI feature stream) absolute and relative to the baseline results (as in Table A.4). See Chapter 8 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	8.10%	9.13%	9.80%	8.85%
Clean	12.00%	14.07%	13.07%	13.04%
Average	10.05%	11.60%	11.43%	10.95%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	37.55%	37.79%	38.51%	37.84%
Clean	76.00%	76.16%	70.62%	74.99%
Average	56.77%	56.97%	54.57%	56.41%

Table A.20: Aurora 3 (SpeechDat-car): Summary of recognition performance of the Tandem G1d system as described above in Tab. A.19.

Aurora 3 Reference Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	7.26%	7.06%	8.80%	12.72%	8.96%
Mid (x35%)	19.49%	16.69%	18.96%	32.68%	21.96%
High (x25%)	59.47%	48.45%	26.83%	60.63%	48.85%
Overall	24.59%	20.78%	16.86%	31.68%	23.48%

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	2.47%	2.10%	5.23%	6.41%	4.05%
Mid (x35%)	8.82%	4.95%	11.71%	19.01%	11.12%
High (x25%)	13.60%	13.53%	11.79%	20.07%	14.75%
Overall	7.47%	5.96%	9.14%	14.24%	9.20%

.

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	65.98%	70.25%	40.57%	49.61%	56.60%
Mid (x35%)	54.75%	70.34%	38.24%	41.83%	51.29%
High (x25%)	77.13%	72.07%	56.06%	66.90%	68.04%
Overall	64.84%	70.74%	43.62%	51.21%	57.60%

Table A.21: Absolute recognition results for the Tandem G2d system (Gabor set optimized on TIMIT phoneme inter-/within group discrimination and combined with the Qualcomm-ICSI-OGI feature stream) on Aurora 2 (TIDigits) and 3 (SpeechDat-car). See Chapter 8 for further description.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.26	99.15	99.16	99.54	99.28	99.26	99.15	99.16	99.54	99.28	99.42	99.15	99.29	99.28
20 dB	99.08	98.97	98.78	99.14	98.99	98.43	98.82	98.84	98.98	98.77	98.80	98.58	98.69	98.84
15 dB	98.53	98.46	98.45	98.18	98.41	97.94	98.22	98.42	98.40	98.25	98.50	97.70	98.10	98.28
10 dB	96.71	96.31	97.02	96.36	96.60	95.12	95.86	96.57	96.64	96.05	96.35	95.89	96.12	96.28
5 dB	92.35	90.15	93.08	90.10	91.42	87.53	89.66	91.50	90.31	89.75	90.88	88.88	89.88	90.44
0 dB	76.17	68.14	78.41	74.36	74.27	64.14	74.00	75.10	74.14	71.85	72.89	69.04	70.97	72.64
-5dB	42.31	27.81	38.80	43.41	38.08	26.22	39.23	38.20	39.23	35.72	35.19	33.01	34.10	36.34
Average	92.57	90.41	93.15	91.63	91.94	88.63	91.31	92.09	91.69	90.93	91.48	90.02	90.75	91.30

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.60	99.40	99.46	99.51	99.49	99.60	99.40	99.46	99.51	99.49	99.57	99.46	99.52	99.50
20 dB	99.02	98.76	99.08	98.86	98.93	98.53	98.70	98.87	98.89	98.75	98.77	98.46	98.62	98.79
15 dB	97.45	97.34	98.21	97.41	97.60	96.25	97.82	97.82	97.84	97.43	98.00	97.19	97.60	97.53
10 dB	94.69	92.74	96.24	94.51	94.55	90.82	94.29	94.36	95.03	93.63	94.69	93.53	94.11	94.09
5 dB	87.44	78.72	89.35	85.71	85.31	75.74	85.07	84.01	85.53	82.59	86.15	83.16	84.66	84.09
0 dB	66.04	46.25	66.95	63.65	60.72	44.92	61.79	56.70	63.59	56.75	61.07	58.88	59.98	58.98
-5dB	30.09	12.88	26.99	29.59	24.89	12.25	28.13	22.61	29.01	23.00	25.08	26.14	25.61	24.28
Average	88.93	82.76	89.97	88.03	87.42	81.25	87.53	86.35	88.18	85.83	87.74	86.24	86.99	86.70

Aurora3 SpeechDat-Car. Accuracy															
Finnish			Spanish			German			Danish			Average			
wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm
97.27	91.52	88.37	97.95	94.56	90.53	95.09	88.58	88.30	93.62	80.86	78.28	95.98	88.88	86.37

Table A.22: Recognition performance of the Tandem G2d system (Gabor set optimized on TIMIT phoneme inter-/within group discrimination and combined with the Qualcomm-ICSI-OGI feature stream) on Aurora 2 (TIDigits) and 3 (SpeechDat-car) relative to the baseline results (as in Table A.4). See Chapter 8 for further description.

Aurora2 TIDigits. Relative Improvement. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	47.52%	36.09%	41.26%	60.68%	46.39%	47.52%	36.09%	41.26%	60.68%	46.39%	57.97%	36.09%	47.03%	46.52%
20 dB	57.80%	50.00%	30.68%	66.01%	51.12%	30.84%	50.63%	51.46%	56.41%	47.33%	48.50%	41.32%	44.91%	48.36%
15 dB	56.12%	40.08%	32.61%	41.67%	42.62%	46.35%	46.55%	53.53%	57.10%	50.88%	57.14%	37.67%	47.41%	46.88%
10 dB	41.46%	18.54%	21.99%	38.20%	30.05%	30.88%	19.46%	45.47%	44.74%	35.13%	40.84%	32.40%	36.62%	33.40%
5 dB	30.39%	16.45%	44.51%	20.16%	27.88%	16.59%	22.95%	31.84%	34.70%	26.52%	46.00%	29.80%	37.90%	29.34%
0 dB	25.88%	13.47%	52.96%	29.35%	30.41%	8.33%	29.62%	26.18%	38.33%	25.61%	49.60%	29.07%	39.34%	30.28%
-5dB	21.45%	0.66%	23.29%	25.90%	17.82%	-1.22%	15.99%	11.83%	22.34%	12.23%	19.77%	11.00%	15.39%	15.10%
Average	42.33%	27.71%	36.55%	39.08%	36.42%	26.60%	33.84%	41.70%	46.26%	37.10%	48.42%	34.05%	41.24%	37.65%

Aurora2 TIDigits. Relative Improvement. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	63.96%	38.14%	43.16%	33.78%	44.76%	63.96%	38.14%	43.16%	33.78%	44.76%	48.19%	40.66%	44.43%	44.70%
20 dB	69.85%	86.89%	68.49%	70.00%	73.81%	85.09%	68.60%	88.76%	78.69%	80.29%	81.45%	68.38%	74.91%	76.62%
15 dB	69.89%	90.44%	84.37%	74.02%	79.68%	85.28%	81.60%	91.67%	88.49%	86.76%	85.68%	74.24%	79.96%	82.57%
10 dB	78.30%	86.14%	89.69%	80.19%	83.58%	80.92%	83.18%	88.88%	88.91%	85.47%	81.08%	74.09%	77.58%	83.14%
5 dB	76.15%	72.40%	84.62%	76.54%	77.43%	66.86%	76.57%	78.79%	80.72%	75.73%	71.95%	65.93%	68.94%	75.05%
0 dB	56.21%	43.10%	62.99%	57.61%	54.98%	40.70%	53.77%	51.62%	59.77%	51.46%	48.42%	46.15%	47.28%	52.03%
-5dB	21.76%	12.78%	21.64%	24.41%	20.15%	11.41%	21.35%	18.30%	24.37%	18.86%	13.98%	16.83%	15.41%	18.68%
Average	70.08%	75.79%	78.03%	71.67%	73.89%	71.77%	72.74%	79.94%	79.31%	75.94%	73.71%	65.76%	69.74%	73.88%

Aurora3 SpeechDat-Car. Relative Improvement.															
Finnish			Spanish			German			Danish			Average		
wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm
62.40%	56.49%	80.44%	70.96%	67.41%	80.45%	44.20%	39.77%	56.39%	49.84%	41.43%	64.18%	56.85%	51.27%	70.37%

Table A.23: Aurora 2 (TIDigits): Summary of recognition performance of the Tandem G2d system (Gabor set optimized on TIMIT phoneme inter-/within group discrimination and combined with the Qualcomm-ICSI-OGI feature stream) absolute and relative to the baseline results (as in Table A.4). See Chapter 8 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	8.06%	9.07%	9.25%	8.70%
Clean	12.58%	14.17%	13.01%	13.30%
Average	10.32%	11.62%	11.13%	11.00%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	36.42%	37.10%	41.24%	37.65%
Clean	73.89%	75.94%	69.74%	73.88%
Average	55.16%	56.52%	55.49%	55.77%

Table A.24: Aurora 3 (SpeechDat-car): Summary of recognition performance of the Tandem G2d system as described above in Tab. A.23.

Aurora 3 Reference Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	7.26%	7.06%	8.80%	12.72%	8.96%
Mid (x35%)	19.49%	16.69%	18.96%	32.68%	21.96%
High (x25%)	59.47%	48.45%	26.83%	60.63%	48.85%
Overall	24.59%	20.78%	16.86%	31.68%	23.48%

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	2.73%	2.05%	4.91%	6.38%	4.02%
Mid (x35%)	8.48%	5.44%	11.42%	19.14%	11.12%
High (x25%)	11.63%	9.47%	11.70%	21.72%	13.63%
Overall	6.97%	5.09%	8.89%	14.68%	8.91%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	62.40%	70.96%	44.20%	49.84%	56.85%
Mid (x35%)	56.49%	67.41%	39.77%	41.43%	51.27%
High (x25%)	80.44%	80.45%	56.39%	64.18%	70.37%
Overall	64.84%	72.09%	45.70%	50.48%	58.28%

Table A.25: Absolute recognition results for the Tandem G3d system (Gabor set optimized on zifkom German digits and combined with the Qualcomm-ICSI-OGI feature stream) on Aurora 2 (TIDigits) and 3 (SpeechDat-car). See Chapter 8 for further description.

Aurora2 TIDigits. Accuracy. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.14	99.33	99.22	99.32	99.25	99.14	99.33	99.22	99.32	99.25	99.11	99.24	99.18	99.24
20 dB	98.93	98.85	98.78	99.01	98.89	98.56	98.40	98.66	99.04	98.67	99.08	98.34	98.71	98.77
15 dB	98.31	98.19	98.39	98.27	98.29	98.04	97.82	98.33	98.27	98.12	98.50	97.61	98.06	98.17
10 dB	96.19	96.31	96.75	96.27	96.38	95.49	95.86	96.72	96.33	96.10	96.13	95.41	95.77	96.15
5 dB	91.53	90.21	93.38	90.47	91.40	88.15	90.54	91.83	90.96	90.37	90.76	89.18	89.97	90.70
0 dB	75.68	69.32	79.30	74.05	74.59	66.44	74.61	75.54	75.07	72.92	72.67	69.46	71.07	73.21
-5dB	41.17	29.59	40.47	40.91	38.04	30.40	39.93	39.34	39.14	37.20	34.85	33.98	34.42	36.98
Average	92.13	90.58	93.32	91.61	91.91	89.34	91.45	92.22	91.93	91.23	91.43	90.00	90.71	91.40

Aurora2 TIDigits. Accuracy. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.36	99.40	99.34	99.48	99.40	99.36	99.40	99.34	99.48	99.40	99.39	99.49	99.44	99.40
20 dB	98.93	98.67	98.99	98.83	98.86	98.59	98.70	98.78	98.67	98.69	99.02	98.58	98.80	98.78
15 dB	98.00	97.43	98.27	97.53	97.81	96.44	97.76	97.97	97.41	97.40	97.85	97.16	97.51	97.58
10 dB	94.60	92.20	96.45	94.57	94.46	91.28	94.53	94.60	95.22	93.91	94.84	94.11	94.48	94.24
5 dB	87.78	80.02	90.87	85.65	86.08	77.16	86.03	85.71	85.96	83.72	87.75	85.13	86.44	85.21
0 dB	70.22	50.24	71.31	65.07	64.21	49.46	66.23	61.94	64.92	60.64	65.98	62.20	64.09	62.76
-5dB	36.66	14.93	32.09	33.48	29.29	16.49	32.43	27.59	32.78	27.32	31.62	29.24	30.43	28.73
Average	89.91	83.71	91.18	88.33	88.28	82.59	88.65	87.80	88.44	86.87	89.09	87.44	88.26	87.71

Aurora3 SpeechDat-Car. Accuracy															
Finnish			Spanish			German			Danish			Average			
wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm
97.27	89.19	89.75	97.86	95.86	91.82	94.57	88.29	88.39	93.59	80.99	78.61	95.82	88.58	87.14

Table A.26: Recognition performance of the Tandem G3d system (Gabor set optimized on zifkom German digits and combined with the Qualcomm-ICSI-OGI feature stream) on Aurora 2 (TIDigits) and 3 (SpeechDat-car) relative to the baseline results (as in Table A.4). See Chapter 8 for further description.

Aurora2 TIDigits. Relative Improvement. Multicondition training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	39.01%	49.62%	45.45%	41.88%	43.99%	39.01%	49.62%	45.45%	41.88%	43.99%	35.51%	42.86%	39.18%	43.03%
20 dB	50.92%	44.17%	30.68%	60.87%	46.66%	36.56%	33.05%	43.93%	58.97%	43.13%	60.52%	31.40%	45.96%	45.11%
15 dB	49.55%	29.57%	30.00%	44.55%	38.42%	48.96%	34.53%	50.88%	53.62%	47.00%	57.14%	35.23%	46.19%	43.40%
10 dB	32.21%	18.54%	14.92%	36.67%	25.59%	36.12%	19.46%	47.85%	39.64%	35.77%	37.28%	24.51%	30.89%	30.72%
5 dB	22.93%	16.96%	46.91%	23.15%	27.49%	20.74%	29.51%	34.48%	39.08%	30.95%	45.29%	31.69%	38.49%	31.07%
0 dB	24.35%	16.68%	54.90%	28.49%	31.11%	14.21%	31.27%	27.48%	40.54%	28.38%	49.19%	30.03%	39.61%	31.72%
-5dB	19.89%	3.11%	25.38%	22.63%	17.75%	4.51%	16.96%	13.45%	22.22%	14.29%	19.35%	12.29%	15.82%	15.98%
Average	35.99%	25.19%	35.48%	38.75%	33.85%	31.32%	29.56%	40.93%	46.37%	37.05%	49.88%	30.57%	40.23%	36.40%

Aurora2 TIDigits. Relative Improvement. Clean training. multicondition testing														
	A					B					C			
	Subway	Babble	Car	Exhibit.	Average	Rest.	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	42.34%	38.14%	30.53%	29.73%	35.19%	42.34%	38.14%	30.53%	29.73%	35.19%	26.51%	43.96%	35.23%	35.19%
20 dB	67.08%	85.94%	65.41%	69.21%	71.91%	85.70%	68.60%	87.86%	74.47%	79.16%	85.22%	70.84%	78.03%	76.03%
15 dB	76.39%	90.76%	84.89%	75.23%	81.82%	86.03%	81.10%	92.24%	86.19%	86.39%	84.61%	73.97%	79.29%	83.14%
10 dB	77.93%	85.11%	90.27%	80.40%	83.43%	81.87%	83.89%	89.36%	89.33%	86.11%	81.61%	76.41%	79.01%	83.62%
5 dB	76.79%	74.08%	86.82%	76.44%	78.53%	68.80%	78.08%	81.05%	81.29%	77.30%	75.19%	69.92%	72.55%	76.85%
0 dB	61.60%	47.33%	67.87%	59.27%	59.02%	45.59%	59.14%	57.47%	61.24%	55.86%	54.92%	50.50%	52.71%	56.49%
-5dB	29.11%	14.83%	27.11%	28.59%	24.91%	15.69%	26.06%	23.55%	28.38%	23.42%	21.49%	20.32%	20.91%	23.51%
Average	71.96%	76.64%	79.05%	72.11%	74.94%	73.60%	74.16%	81.60%	78.50%	76.96%	76.31%	68.33%	72.32%	75.23%

Aurora3 SpeechDat-Car. Relative Improvement.															
Finnish			Spanish			German			Danish			Average			
wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm	wm	mm	hm
62.40%	44.54%	82.76%	69.69%	75.19%	83.12%	38.30%	38.24%	56.73%	49.61%	41.83%	64.72%	55.00%	49.95%	71.83%

Table A.27: Aurora 2 (TIDigits): Summary of recognition performance of the Tandem G3d system (Gabor set optimized on zifkom German digits and combined with the Qualcomm-ICSI-OGI feature stream) absolute and relative to the baseline results (as in Table A.4). See Chapter 8 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	8.09%	8.77%	9.29%	8.60%
Clean	11.72%	13.13%	11.74%	12.29%
Average	9.90%	10.95%	10.51%	10.44%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	33.85%	37.05%	40.23%	36.40%
Clean	74.94%	76.96%	72.32%	75.23%
Average	54.40%	57.00%	56.27%	55.82%

Table A.28: Aurora 3 (SpeechDat-car): Summary of recognition performance of the Tandem G3d system as described above in Tab. A.27.

Aurora 3 Reference Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	7.26%	7.06%	8.80%	12.72%	8.96%
Mid (x35%)	19.49%	16.69%	18.96%	32.68%	21.96%
High (x25%)	59.47%	48.45%	26.83%	60.63%	48.85%
Overall	24.59%	20.78%	16.86%	31.68%	23.48%

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	2.73%	2.14%	5.43%	6.41%	4.18%
Mid (x35%)	10.81%	4.14%	11.71%	19.01%	11.42%
High (x25%)	10.25%	8.18%	11.61%	21.39%	12.86%
Overall	7.44%	4.35%	9.17%	14.57%	8.88%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	62.40%	69.69%	38.30%	49.61%	55.00%
Mid (x35%)	44.54%	75.19%	38.24%	41.83%	49.95%
High (x25%)	82.76%	83.12%	56.73%	64.72%	71.83%
Overall	61.24%	74.97%	42.88%	50.66%	57.44%

Table A.29: Aurora 2 (TIDigits): Summary of recognition performance of the melspec Tandem system R2d (nine frames of context) combined with the Qualcomm-ICSI-OGI feature stream absolute and relative to the baseline results (as in Table A.4). See Chapter 8 for further description.

Aurora 2 Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	11.93%	12.78%	15.44%	12.97%
Clean	41.26%	46.60%	34.00%	41.94%
Average	26.59%	29.69%	24.72%	27.46%

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	8.44%	9.40%	10.21%	9.18%
Clean	13.26%	15.63%	12.29%	14.01%
Average	10.85%	12.52%	11.25%	11.60%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	33.69%	34.92%	35.52%	34.55%
Clean	72.42%	73.45%	69.72%	72.29%
Average	53.06%	54.19%	52.62%	53.42%

Table A.30: Aurora 3 (SpeechDat-car): Summary for R2d as described above in Tab. A.29.

Aurora 3 Reference Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	7.26%	7.06%	8.80%	12.72%	8.96%
Mid (x35%)	19.49%	16.69%	18.96%	32.68%	21.96%
High (x25%)	59.47%	48.45%	26.83%	60.63%	48.85%
Overall	24.59%	20.78%	16.86%	31.68%	23.48%

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	2.83%	2.21%	5.05%	6.72%	4.20%
Mid (x35%)	10.88%	4.12%	11.49%	20.70%	11.80%
High (x25%)	11.80%	10.44%	10.78%	21.64%	13.67%
Overall	7.89%	4.94%	8.74%	15.34%	9.23%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	61.02%	68.70%	42.61%	47.17%	54.87%
Mid (x35%)	44.18%	75.31%	39.40%	36.66%	48.89%
High (x25%)	80.16%	78.45%	59.82%	64.31%	70.68%
Overall	59.91%	73.45%	45.79%	47.78%	56.73%

List of Figures

1.1 Schematic overview of the processing steps in ASR 11
1.2 Spectral, temporal and spectro-temporal processing. 16

2.1 Processing stages of the auditory model (PEMO). 25
2.2 Modulation transfer function of PEMO. 27
2.3 Recognition rates in CCITT noise and Sotscheck noise . . 30
2.4 Examples for PEMO processing of speech. 31
2.5 Setup of isolated digit recognition experiment. 32
2.6 Digit recognition rates in CCITT noise applying Ephraim-Malah speech enhancement. 34
2.7 Recognition rates for different types of noise. 36
2.8 Setup of binaural isolated digit recognition experiment. . . 37
2.9 Recognition rates in CCITT noise as function of SNR in anechoic condition. 40
2.10 Recognition rates in CCITT noise as function of SNR in reverberant condition. 41
2.11 Recognition rates in anechoic conditions - comparison monaural and binaural processing. 43
2.12 Recognition rates in reverberant conditions - comparison monaural and binaural processing. 45

3.1 Scheme of the PEMO/FFNN recognition system. 54
3.2 Word error rates in depending on the number of secondary features. 56
3.3 Word recognition scores for FFNN and LRNN. 58

4.1 Spectro-temporal representation of phoneme /d/. 65
4.2 Spectro-temporal representation of phoneme /t/. 66
4.3 Histogram of 500 features with highest Fisher Score for short vowels. 71
4.4 Histogram of 500 features with highest Fisher Score for unvoiced plosives. 72

5.1 Transfer functions of the modulation filters. 80
5.2 Schematic overview of sigma-pi cell calculation. 82
5.3 Histogram of segmental SNR. 88
5.4 Processing steps in the SNR estimator. 89
5.5 Percentiles of estimation error 90
5.6 Error distribution over frequency channels 91
5.7 Example of PEMO primary features 94
5.8 Classification error dependency on the number of secondary features - linear network. 95
5.9 Classification error dependency on the number of secondary features - multi-layer perceptron. 96
5.10 Classification error depending on the segment length. . . . 97
5.11 Classification error dependency on modulation frequency. . 99
5.12 Example for SNR estimation 101

6.1 Example of a primary feature matrix 105
6.2 Parameter sketch for a sigma-pi cell with two windows . . 108
6.3 Example of a sigma-pi cell and resulting feature values. . . 109
6.4 Example of a 2D Gabor function and resulting feature values. 113

7.1 Example of a one-dimensional complex Gabor function . . 125
7.2 Example of Gabor complex filtering. 127
7.3 Distribution of Gabor types for different target labels. . . . 128
7.4 Distribution of temporal modulation frequency. 129
7.5 Distribution of spectral modulation frequency. 131
7.6 Sketch of the Gabor Tandem recognition system. 132

7.7 Relative improvement over the Aurora 2 baseline 135

8.1 Combination of the Gabor Tandem system and the Qualcomm-ICSI-OGI proposal. 141
8.2 Results for Aurora 2. 145
8.3 Overview of results for Aurora 2 and 3 corpora. 146

A.1 Distribution of the three modes of Gabor features in the optimized sets. 164
A.2 Distribution of the four types of Gabor features in the optimized sets for within group discrimination. 165
A.3 Overview of Gabor set G1. 166
A.4 Overview of Gabor set G2. 168
A.5 Overview of Gabor set G3. 170

List of Tables

2.1 Recognition rate on clean test data. 35
2.2 Recognition rate for different reverberant environments. . . 42

4.1 Recognition scores on test data and the corresponding optimal integration time. 69
4.2 Results of the Fisher score ranking. 70
4.3 Statistics of the 100 features with highest Fisher score. . . 70

5.1 Frequency of different noise types in the training set. . . . 78
5.2 SNR estimation error for different noise categories. 91
5.3 Estimation error dependency on primary feature extraction. 93
5.4 RMS error depending on the type of sigma-pi cells. 95
5.5 Estimated computational effort 98

6.1 Word error rates - comparison of the three feature types. . 117
6.2 Word error rates - comparison with Aurora 2 baseline. . . 119

7.1 WER and improvement over the Aurora 2 baseline 134

8.1 Performance of different front ends for Aurora 2. 143
8.2 Performance of different front ends for Aurora 2 and 3. . . 144

A.1 Table of parameters for Gabor set G1. 167
A.2 Table of parameters for Gabor set G2. 169
A.3 Table of parameters for Gabor set G3. 171
A.4 Baseline front end performance on Aurora 2 and 3. 172

A.5 Aurora 2 absolute recognition results for the Tandem G1-R1-P system. 173
A.6 Aurora 2 relative recognition results for the Tandem G1-R1-P system. 174
A.7 Aurora 2 absolute recognition results for the Tandem G3-R1-P system. 175
A.8 Aurora 2 relative recognition results for the Tandem G3-R1-P system. 176
A.9 Aurora 2 summary of recognition performance of the Tandem G1-R1-P system. 177
A.10 Aurora 2 summary of recognition performance of the Tandem G3-R1-P system. 177
A.11 Aurora 2 absolute recognition results for the Tandem G1-D system. 178
A.12 Aurora 2 relative recognition results for the Tandem G1-D system. 179
A.13 Aurora 2 absolute recognition results for the Tandem G1-R0-Q system. 180
A.14 Aurora 2 relative recognition results for the Tandem G1-R0-Q system 181
A.15 Aurora 2 summary of recognition performance of the Tandem G1-D system. 182
A.16 Aurora 2 summary of recognition performance of the Tandem G1-R0-Q system. 182
A.17 Absolute recognition results for the Tandem G1d system. . 183
A.18 Relative recognition performance of the Tandem G1d system. 184
A.19 Aurora 2 summary of recognition performance of the Tandem G1d system. 185
A.20 Aurora 3 summary of recognition performance of the Tandem G1d system. 185
A.21 Absolute recognition results for the Tandem G2d system. . 186
A.22 Relative recognition performance of the Tandem G2d system. 187
A.23 Aurora 2 summary of recognition performance of the Tandem G2d system. 188

A.24 Aurora 3 summary of recognition performance of the Tandem G2d system. 188
A.25 Absolute recognition results for the Tandem G3d system. . 189
A.26 Relative recognition performance of the Tandem G3d system.190
A.27 Aurora 2 summary of recognition performance of the Tandem G3d system. 191
A.28 Aurora 3 summary of recognition performance of the Tandem G3d system. 191
A.29 Aurora 2 summary of recognition performance of the melspec Tandem system. 192
A.30 Aurora 3 summary of recognition performance of the melspec Tandem system. 192

Danksagung

Zu allererst möchte ich mich bei Prof. Dr. Dr. Birger Kollmeier herzlich für die mir zuteil gewordene Unterstützung bedanken. Durch seinen unermüdlichen Einsatz hat er die Rahmenbedingungen für diese Promotion geschaffen und durch seine Erfahrung, seine Ermutigung, sein Vertrauen und starken Rückhalt selbige in vielerlei Hinsicht geprägt.
Ein großer Dank geht ebenfalls an Dr. Volker Hohmann für die fachkundige und freundschaftliche Zusammenarbeit, die mir viel Freude bereitet hat.
Meine Erinnerung und mein Respekt gilt Dr. Tino Gramß, in dessen Fußstapfen ich durch die Wahl des Themas zwangsläufig geraten bin, ohne sie jedoch nur annähernd ausfüllen zu können.
Dr. Jürgen Tchorz möchte ich für die langjährige gute und oft sehr humorvolle Zusammenarbeit danken. Dr. Mark Mazinzik, Dr. Jörn Anemüller und Dr. Thomas Wittkop haben meine Arbeit ebenfalls durch ihre Kooperation inhaltlich und menschlich bereichert.
Prof. Dr. Volker Mellert danke ich für die freundliche Übernahme des Korreferats.
Mein Dank geht insbesondere auch an PD Dr. Christian Kaernbach für die spannenden Tage in Leipzig und viele Anregungen.
Den ehemaligen und jetzigen Mitgliedern der Arbeitsgruppe Medizinische Physik möchte ich für ein wirklich angenehmes Arbeitsklima, vielfältige Diskussionen, Anregungen und all die kleinen Hilfestellungen danken. Zu Dank verpflichtet bin ich auch meinen Administrator-Kollegen Johannes Nix, Dr. Jörn Anemüller, Dr. Oliver Wegner und vielen anderen, die durch ihre hohe Frustrationstoleranz so viel Last von anderen nehmen und immer gut auf die beiden 'Höllenhunde' Castor und Pollux aufgepasst haben.
Meinen 'Mitbewohnern' Rainer Huber und Dr. Jörn Otten möchte ich herzlich für das gute Raumklima, viele Anregungen und Diskussionen danken, die mir immer angenehm in Erinnerung bleiben werden.
Meiner Frau Sabine Lattemann danke ich für so vieles, speziell jedoch für die unbezahlbare LaTeX'nische Unterstützung!

Acknowledgements

I really want to express my gratitude to Prof. Dr. Nelson Morgan and Dr. Steven Greenberg who made my stay at ICSI in 2001/2002 possible, successful and enjoyable. Their competent advice, strong interest in new ideas and warm hospitality made these six months a very memorable experience.
In addition, I would like to thank Prof. Dr. Hynek Hermansky for the few but very valuable discussions we had in Berkeley, for looking into tiny details, really getting the grip on my ideas and giving me good advice.
I am indebted to David Gelbart, not only for the good coorperation, technical support, many many questions and discussion, but also for being a great room mate.
Furthermore I would like to thank all members of ICSI, especially of the speech group and most prominently Dr. Stéphane Dupont and Barry Yue Chen for the good time and much support.

About the author

Michael Kleinschmidt received his BSc Physics (Physik Vordiplom) at *Georg-August Universität Göttingen* in October 1994. He continued his studies at *Victoria University of Wellington*, New Zealand, where he was awarded the Diploma in Applied Science (Meteorology) in March 1996. Having returned to Germany, he finished his MSc Physics (Physik Diplom) at *Carl-von-Ossietzky Universität Oldenburg* in March 1999.
The author then stayed at Universität Oldenburg as a PhD student and staff member for several funded projects about noise reduction and automatic speech recognition in the research group of Prof. Birger Kollmeier. He also became an associated member of the Graduate School 'Psychoakustik' and the European Graduate School for 'Neurosensory Science, Systems and Application'. While working on this dissertation, Michael also helped supervising several Masters students and gave tutorials in the lab courses 'Digital Signal Processing' and 'Computational Physics'. Between October 2001 and March 2002, he worked as a visiting scientist at the *International Computer Science Institute, Berkeley*, California.
This PhD thesis is the outcome of three years of intense research, culminating in the successful defense of this work on September 5, 2002, at *Carl-von-Ossietzky Universität Oldenburg*.