Marc René Schädler

Robust automatic speech recognition and modeling of auditory discrimination experiments with auditory spectro-temporal features

BIS-Verlag der Carl von Ossietzky Universität Oldenburg

Oldenburg, 2016

Verlag / Druck / Vertrieb

BIS-Verlag
der Carl von Ossietzky Universität Oldenburg
Postfach 2541
26015 Oldenburg
E-Mail: bisverlag@uni-oldenburg.de

ISBN 978-3-8142-2333-9

Preface

Automatic speech recognition (ASR) these days seems to be a major technology mastered by big companies with large research teams that provide speech recognition services for millions of people and applications – so why should a PhD thesis in physics be concerned with this technology? Admittedly, using the limited resources of a single brain, the strict time limit of a few years and the limited computer resources of a university environment, typically associated with a dissertation, mean an uphill battle when trying to compete with Google and similar giants – how big are the chances to achieve any significant progress in this area?

If the reader wants a response to these questions, she or he should read this dissertation: In the end, it is the brain that counts, a set of good ideas, and the stimulating research environment that make the difference. Marc René Schädler has succeeded to advance the field of automatic speech recognition for acoustically difficult conditions in noise by implementing knowledge from neurophysiology and psychophysics – thus providing ears to the computer!

The biologically inspired Gabor feature sets proposed by him are shown to work better for a number of conditions than the standard speech recognition technology used so far. Moreover, Marc René demonstrated the advantage of so-called separable Gabor feature filter banks (that are composed of a pure spectral and a pure temporal processing part) in comparison to the complete two-dimensional time-frequency Gabor feature representation. This allows to speculate that spectral resolution and temporal resolution in the auditory system are not as coupled as has been thought before – please read yourself to find out!

But Marc René would not be himself if he did not provide significant progress to other fields as well that are classically separated from automatic speech recognition: By applying his ASR technology to predict human speech recognition in a specialized task (the so-called Matrix sentence test using syntactically fixed, but semantically unpredictable sentences) a surprisingly well-working prediction of human speech recognition (HSR) becomes possible. It is solely based on the available speech material, the recognition task employed and some general ASR principles

– and produces a prediction accuracy for speech in stationary noise, which was totally unexpected! Please find out yourself how Marc René's new method clearly outperforms the "work horse" of speech intelligibility modeling on which generations of researchers have already elaborated for decades, i.e., the speech intelligibility index (SII).

Moreover, Marc René carries this approach further towards predicting classical psychoacoustic experiments using the same ASR front end as developed before and a standard speech recognizer back end – an approach that seems as unorthodox as to predict the result of expert wine tasting using methods from quality control of fabricating sausages – but it works! It is this non-standard, non-mainstream, unorthodox thinking that enables Marc René to master not only the field of automatic speech recognition, but also to excel in speech intelligibility prediction and in psychoacoustic modeling – always using the same approach which he then terms the "FADE framework". This approach has the great potential of connecting these three classically diverse fields by employing a common, unifying modeling method – please yourself to find out how Marc René achieved this great task!

Besides all the work with machines and implementing physiology- and psychology motivated models it should not be forgotten that this work is just one aspect of Marc René's great personality and life as a researcher, colleague and caring family father: His inspiring, open and friendly way of communicating with colleagues and friends, his overwhelming charm to motivate coworkers to collaborate with him, and his preparedness to take care of nearly all social affairs in the department (from the weekly "Medi-breakfast" to organizing research group meetings and student exchange visits) have been very valuable components of his dissertation project. It has been a great pleasure to work with him and I have no doubt that Marc René Schädler has a great career as a scientist ahead of him – please read yourself to get convinced!

Birger Kollmeier, November 2015

Abstract

Automatic speech recognition (ASR) systems still do not perform as well as human listeners under realistic listening conditions. The vast variability of speech signals in the world is too much for any of the current approaches. Even under selected "lab conditions" no current recognition system reaches the performance of listeners with normal hearing. The unmatched ability of humans to understand speech in the most difficult conditions originates from the superior properties of their auditory system. Hence, this thesis focuses on integrating knowledge about the auditory system into ASR systems in order to improve their capacity to cope with realistic listening conditions, which is also referred to as robustness. A long-term strategy to achieve human-like performance on ASR tasks, which includes validating the correct implementation of auditory principles into ASR systems is proposed, where the ASR systems are considered to be "auditory models" and are compared to the human perception of sound.

In the first part of this work, the physiologically-inspired extraction of spectro-temporal modulation patterns was integrated into a standard ASR system to successfully improve its robustness. These patterns were extracted with a bank of two-dimensional Gabor filters (GBFB) which perform a joint spectro-temporal processing. In addition, it was shown that the joint spectro-temporal filter processing can be replaced by a separate spectral and temporal one which extracts separable Gabor filter bank (SGBFB) features and further increases the robustness of a standard ASR system. In the second part of this work, ASR systems were employed to simulate auditory discrimination experiments in order to test their "auditory fidelity". The employed experiments were tone-in-noise detection experiments and speech intelligibility tests in stationary and fluctuating noise conditions. The performance with several feature extraction algorithms, including traditional ASR features and an effective model of the human auditory signal processing, were compared to empirical data. The standard ASR system using SGBFB features was found to provide the most suitable model of human performance across the considered experiments. Further, the spectral modulation processing was found to be crucial to recognize speech in fluctuating noise.

The fact that SGBFB features faithfully implement some basic auditory principles, i.e., an ASR system using SGBFB features obtained human-

like performance in the corresponding experiments, and improve the robustness of a standard ASR system could be interpreted as a faint hint that spectral and temporal modulations in the human auditory system might be separately processed. In future work, it would be interesting to investigate which auditory principles are essential to achieve human-like performance with ASR systems in increasingly more complex (speech) recognition tasks.

Zusammenfassung

Automatische Spracherkennungssysteme (*engl.* automatic speech recognition (ASR) systems) erreichen unter realistischen Kommunikationsbedingungen noch immer nicht die Spracherkennungsleistung menschlicher Hörer. Die immense Variabilität von Sprachsignalen auf der Welt überfordert alle aktuellen Ansätze zur automatischen Spracherkennung. Selbst unter ausgewählten „Laborbedingungen“ erreichen heutige ASR-Systeme nicht die Leistung normal hörender Menschen. Die Fähigkeit gesprochene Sprache unter den schwierigsten akustischen Bedingungen zu verstehen, verdanken wir den besonderen Eigenschaften unseres auditorischen Systems. Aus diesem Grund bildet die Integration von Wissen über das menschliche auditorische System in ASR-Systeme, mit dem Ziel der Verbesserung deren Erkennungsleistung in realistischen Hörsituationen (auch Robustheit genannt), den Kern dieser Arbeit. Eine langfristig ausgelegte Strategie, um mit ASR-Systemen menschenähnliche Erkennungsleistung zu erzielen, wird vorgestellt. Der Hauptpunkt der Strategie ist zu überprüfen, ob auditorische Prinzipien wirkungsvoll in ASR-Systeme implementiert wurden, wozu diese als auditorische Modelle betrachtet und in Experimenten mit der menschlichen Schallwahrnehmung verglichen werden.

Im ersten Teil der Arbeit wurde die physiologisch motivierte Extraktion spektro-temporaler Modulationsmuster erfolgreich in ein gängiges ASR-System integriert und dessen Robustheit dadurch verbessert. Diese Muster wurden mit einer Filterbank zweidimensionaler Gaborfilter (GBFB), welche sich durch eine gekoppelte spektro-temporale Signalverarbeitung auszeichnen, berechnet. Zudem wurde gezeigt, dass die gekoppelte spektro-temporale Modulationsverarbeitung der GBFB-Merkmale durch eine entkoppelte (*engl.* separable), rein zeitliche und rein spektrale Modulationsverarbeitung ersetzt werden und die Robustheit eines gängigen ASR-System mit diesen separablen GBFB (SGBFB) Merkmalen sogar noch weiter verbessert werden kann.

Im zweiten Teil der Arbeit wurden auditorische Diskriminationsexperimente mit ASR-Systemen simuliert, um deren „Modellierungstreue“ zu bestimmen. Die eingesetzten Experimente waren „Ton-in-Rauschen“ Detektionsexperimente und Sprachverständlichkeitstests in stationärem und fluktuierendem Störgeräusch. Die Leistung der ASR-Systeme, sowohl mit typischen ASR-Merkmalen als auch mit effektiven Modellen

der menschlichen auditorischen Signalverarbeitung als 'front-end', wurde mit empirischen Daten aus Hörexperimenten verglichen. Unter allen betrachteten Merkmalen waren die SGBFB-Merkmale diejenigen, mit denen die simulierten Ergebnisse am besten mit den empirischen Daten übereinstimmten. In diesem Zusammenhang hat sich in den Simulationen des Sprachverständlichkeitstests in fluktuierendem Störgeräusch die spektrale Modulationsverarbeitung, welche eine frequenzübergreifende Verarbeitung der Signale darstellt, als besonders wichtig herausgestellt.

Der Umstand, dass SGBFB-Merkmale grundlegende auditorische Prinzipien besonders originalgetreu modellieren, was heißt, dass in den entsprechenden Simulationen mit diesen Merkmalen menschenähnliche Ergebnisse erzielt werden, könnte als Anzeichen dafür gedeutet werden, dass spektrale und temporale Modulationen im menschlichen auditorischen System getrennt voneinander verarbeitet werden. Eine interessante Aufgabe für zukünftige Forschungsprojekte wäre es, zu untersuchen, welche auditorischen Prinzipien essentiell sind, um auch in komplexeren (Sprach-) Erkennungsexperimenten menschenähnliche Leistung zu erreichen.

Contents

1 General Introduction

The most natural way of communication between human beings is spoken language. Automatic speech recognition (ASR) is the art of building a machine which is able to transcribe spoken into written language automatically. Ever since humans tried to teach—or in technical terms, to program—machines to recognize *their* language they found that speech recognition is a complex task and that human listeners are incredibly good at it, al least, compared to their automatic counterparts (e.g., Lippmann, 1997; Cooke and Scharenborg, 2008; Meyer et al., 2011b). The reason for this gap between human and machine speech recognition performance *(man-machine gap)* is the vast variability within spoken languages, e.g., gender, dialects, or mood, combined with the—maybe even vaster—variability of the acoustical situations that spoken language is found in, e.g, on the telephone, in crowded places, in a reverberant church hall, or in a cockpit. Hence, the long-term goal of all efforts to improve ASR systems is to reach or surpass human recognition performance. Currently, no universal solution to this problem exists.

A common approach to automatic speech recognition

Most of the common ASR systems consist of, at least, three characteristic components, which are: 1) A corpus of labeled speech data, 2) a signal processing algorithm to extract features, and 3) a classifier. The raw speech data, which may also contain non-speech signals, is usually first processed to extract the features which are relevant for the speech recognition task, e.g., changes in the spectral composition. At the same time the processing often suppresses signal characteristics which do not contain usable speech information, e.g., the "spectral color" of the recording due to the used microphone. During the training stage, the classifier learns the characteristics of the different parts that speech consists of, e.g., phonemes, from the features of the labeled speech data. The trained ASR system can than be used to recognize, i.e., assign labels, to new recordings, which were not part of the training data.

Speech recognition experiments

To compare the performance of different systems, usually recognition experiments are performed. Often, labeled training and testing speech data sets—also called corpora—are defined. The training speech data set is used to train different ASR systems, and the trained ASR systems are then used to assign labels to the test speech data. Comparing the testing data labels and the ASR-assigned labels the performance of the corresponding ASR system can be quantified by the achieved word recognition rate (WRR) or word error rate (WER) in percent.

Robust automatic speech recognition

Humans possess the amazing ability to recognize spoken language of unfamiliar speakers even in acoustically adverse situations, such as the famous cocktail party where a multitude of different sources compete with the target speaker. Hence, humans are said to be *robust* speech recognizers, because they are capable to compensate for the variability in speech signals. To give a more detailed explanation of the shortcomings of ASR systems, their susceptibility to variability in the recordings originating from different sources could be distinguished (Meyer and Kollmeier, 2011a): Variability within the speech signals—also called *intrinsic* variability—, such as dialects, and variability due to the situation—also called *extrinsic* variability—, such as background noise but also reverberation. Of course, the two classes themselves could be in turn subdivided into more specific sources of variability, such as a type of background noise, e.g., stationary vs. modulated, or the speaker's age. Hence, in the context of speech recognition, "robustness" is not clearly defined.

A practical solution to the definition of a robust ASR system is to define a speech recognition experiment in realistic conditions. Then, an ASR system which performs better under these realistic conditions could be considered more robust. Consequently, standardized ASR tasks exist with the objective of comparing the robustness of ASR systems (e.g., Cole et al., 1995; Pearce and Hirsch, 2000; Barker et al., 2013).

Approaches to improve the robustness of ASR systems

In the past, the components of ASR systems were subject to attempts of improving their "robustness", or rather, improving their performance

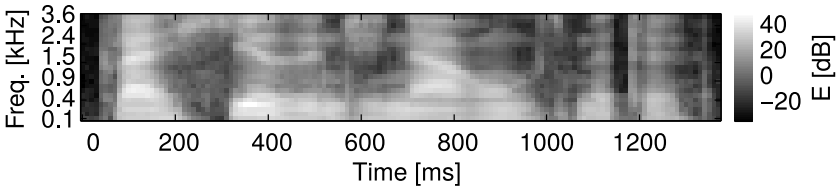

Figure 1.1: *Taken from Schädler and Kollmeier (2012b). Spectro-temporal representation of a recording of read speech. Light areas denote high energy.*

under realistic conditions. Besides technical improvements, mimicking human strategies to recognize speech is a logical approach to achieve human-like recognition performance.

A common, successful strategy that focuses on the signal processing—also *front-end* or *feature extraction*—is to integrate auditory principles into it (e.g, Hermansky, 1990; Tchorz and Kollmeier, 1999; Kleinschmidt and Gelbart, 2002a; Domont et al., 2008; Meyer and Kollmeier, 2011a). These principles were found in, e.g., psycho-acoustical (Stevens, 1957; Zwicker, 1970; Dau et al., 1997) or physiological studies (Depireux et al., 2001; Qiu et al., 2003) and include the limitations of the human auditory system, such as, e.g., a limited spectral resolution or just noticeable differences in sound intensity. The least common denominator for all feature extraction algorithms is a somewhat blurred spectro-temporal representation of the sound signal. It shows the temporal evolution of the frequency-decomposition of a signal over time. Figure 1.1 shows an example of a spectro-temporal representation of a speech signal which is commonly used in ASR systems. Another example of a spectro-temporal representation would be a spectrogram, which can be calculated using a short-term Fourier transform.

In Chapter 2, this work sets out with the aim of improving the robustness of an ASR system and for that purpose introduces an auditory-motivated and physiologically-inspired feature extraction scheme called *Gabor filter bank (GBFB) features.* These features extract spectro-temporal modulation patterns from a spectro-temporal representation, such as the one depicted in Fig. 1.1, and provide high-dimensional ($d = 311$) signal representations compared to traditional features ($d = 39$). The robustness of GBFB features against intrinsic and extrinsic sources

of variability in speech was evaluated and compared to the results of traditional ASR features. In Chapter 3, the effect of normalizing the feature value statistics of GBFB features on the robustness of an ASR system was investigated. While in Chapter 2 and 3 small vocabulary (< 100 words) ASR tasks were targeted, in Chapter 4 an ASR system using GBFB features was evaluated on a medium-size (≈ 5000 words) task. There, it was tested if the high dimensionality of the GBFB signal representation interfered with the more complex recognition task. One observation was that the employed back-end was more complex than in the small-vocabulary task but the differences in performance resulting from the different feature sets were much smaller. Consequently, in Chapter 5, where the hypothesis that the joint spectro-temporal GBFB feature extraction can be divided into a separate spectral and temporal processing was tested, a small-vocabulary recognition task which allowed to compare the ASR systems at different SNRs was employed.

Another approach of integrating biologically-inspired principles into ASR systems, which focuses on the classifier—or *back-end*—instead of the feature extraction, was to extend the systems with artificial neural nets to improve their robustness (e.g., Hermansky et al., 2000; Dahl et al., 2012). Artificial neural networks were inspired by biological neural networks and can be trained as recognizers. A big advantage of neural networks is that they can learn to handle diverse types of input signals, given sufficient training data is provided.

Conflict of objectives between performance and knowledge

Recent developments indicate that increasingly complex artificial neural networks, such as deep neural networks, could eventually replace parts of the feature extraction stage in ASR systems (Castro Martinez et al., 2014) because they possibly can also learn the optimal feature extraction algorithm for a given data set. Unfortunately, the signal processing mechanisms learned by a deep neural network are not easily accessible anymore. This leads to the following questions:

> *Would we be satisfied with super-human ASR performance? Or do we still want to know how speech recognition actually works?*

Besides the severe possible social impact (e.g., mass-surveillance), it seems tempting to bury the burden to deal with the complexity of speech recognition tasks in a black-neural-network-box using massive amounts

of training data. But the gained improvements in performance do not reflect any gained knowledge. We humans can understand speech without a clue about how spoken language works, just like we can ride a bike without understandings its physics. Put another way, one could possibly train a neural network with a gyroscope and attached actuators to ride a bike without learning anything about the physics of riding a bike. Hence, a conflict of objectives can be identified which, on one hand, consists of the desire to build ASR systems which are as robust as a human listeners, and on the other hand, the wish to discover what the different parts of the speech recognition puzzle are and to learn how they interact.

This conflict is directly reflected in the speech recognition experiments by whose means the robustness of ASR systems is defined. On the one hand, there are simple ASR corpora/tasks (Cole et al., 1995; Pearce and Hirsch, 2000; Vincent et al., 2013a) which allow to evaluate the robustness of ASR systems under comparatively controlled, but "unrealistic", conditions, e.g., the recognition of digits or letters at different signal-to-noise ratios (SNRs). On the other hand, there are complex corpora/tasks (Barker et al., 2015, (submitted)) which allow to evaluate the robustness of ASR systems under comparatively realistic, but "uncontrolled", conditions, e.g., the recognition of read Wall street Journal articles in a highly dynamic environment. More complex tasks naturally introduce additional challenges which usually require more complex solutions. The recognition performance on a task reflects how well these challenges were mastered *on average*.

> *The main problem with increasingly complex tasks is that ASR systems still perform worse than human listeners even on simple tasks (e.g., Vincent et al., 2013b).*

While tracking down the origin of the man-machine gap on a simple task poses a difficult problem, it surely seems impossible on a complex task where the recognition results depend on the interactions of a multitude of highly non-linear systems. Hence, realistic and controlled conditions are always a compromise, and the robustness of an ASR system and its specific weaknesses cannot be assessed using a single speech recognition task.

Understanding speech perception *and* improving ASR

> *Before focusing research on increasingly complex speech recognition tasks, the man-machine gap should be closed for the simpler tasks first.*

If one imagines robust ASR as a very high-dimensional parameter optimization problem—and with parameters one could think of every decision which could be taken in the process of building an ASR system—the recognition performance would describe a cliffy and hilly high-dimensional ($d \gg 1$) landscape. The hills would be due to the non-linear interaction between the parameters and the cliffs due some parameter-subspaces being discrete. Even if the parameter space was smooth, we would be stuck in a local optimum trapped by the curse of dimensionality, i.e., there are too many—at least $O(2^d)$—directions that would need to be explored.

The human recognition performance provides evidence that it is a *local* optimum that we are stuck in. Further, considering more complex recognition tasks could be thought of as exponentially inflating the parameter space by adding more dimensions. Even worse, if we improve recognition performance, there is no evidence that the "direction" in which we turned points towards the global optimum, or at least the human local optimum; it might even be the "wrong direction" to go.

A solution to this problem could consist in deflating the parameter space, i.e., taking simpler/more controlled tasks, as much as necessary to find the human local optimum and tracking it afterwards while gradually increasing the task complexity. The outlined path is guided by the human performance and could be seen as taking a short-cut "slipping" past the dimensions. In this work the first steps down this path were taken.

Opportunities along the way

> *Once on track, following the path comes down to meticulously assuring that the recognizer "has ears."*

In a first step, the path, i.e. a recognition experiment in which human and machine performance are on par, needed to be found. Therefore, the key is to employ recognition tasks that allow a direct comparison to the performance of human listeners, preferably tasks for which empirical data already exists. In Chapter 6 hence, speech recognition experiments

simulating a speech intelligibility test in simple, stationary noise conditions were performed. In Chapter 7, the approach was generalized to simulate basic psycho-acoustic experiments to test if automatic (speech) recognition systems could "hear" as well as listeners with normal hearing. This generalized approach can be used to assess the man-machine gap in experiments which belong to the class of auditory discrimination experiments. Further, the speech intelligibility test was simulated in a more challenging, fluctuating noise condition and it was tested if a system with SGBFB features exhibited "more auditory" behavior than recognition systems with other signal representations.

Apparently, the outlined path to more robust ASR systems crosses other disciplines of hearing research, such as the prediction of speech intelligibility and classical psychoacoustic modeling. This might offer the opportunity to comprehend classical modeling of experimental results from a performance-oriented perspective. The most important difference between classical models and the proposed ASR-based approach is that the former only need to perform well in a set of very specific tasks, while the latter eventually must perform all tasks on the way to large vocabulary continuous speech recognition as good as human listeners with normal hearing. Down the road, some of the auditory principles as they are understood today might prove essential while others could prove harmful, and still others may require a recast. Hence, the current thesis might be a first step towards hearing research as viewed from a machine learning perspective which may be denoted as "computational hearing research".

2 Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition

ABSTRACT

In an attempt to increase the robustness of automatic speech recognition (ASR) systems, a feature extraction scheme is proposed that takes spectro-temporal modulation frequencies (MF) into account. This physiologically inspired approach uses a two-dimensional filter bank based on Gabor filters, which limits the redundant information between feature components, and also results in physically interpretable features. Robustness against extrinsic variation (different types of additive noise) and intrinsic variability (arising from changes in speaking rate, effort, and style) is quantified in a series of recognition experiments. The results are compared to reference ASR systems using Mel-frequency cepstral coefficients (MFCCs), MFCCs with cepstral mean subtraction (CMS) and RASTA-PLP features, respectively. Gabor features are shown to be more robust against extrinsic variation than the baseline systems without CMS, with relative improvements of 28% and 16% for two training conditions (using only clean training samples or a mixture of noisy and clean utterances, respectively). When used in a state-of-the-art system, improvements of 14% are observed when spectro-temporal features are concatenated with MFCCs, indicating the complementarity of those feature types. An analysis of the importance of specific MF shows that temporal MF up to 25 Hz and spectral MF up to 0.25 cycles/channel are beneficial for ASR.

This chapter is a reformatted reprint. The original article can be found at http://dx.doi.org/10.1121/1.3699200. Reproduced with permission from "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition", M. R. Schädler, B. T. Meyer, and B. Kollmeier, J. Acoust. Soc. Am. Vol. 131, pp. 4134–4151. Copyright 2012, Acoustical Society of America.

2.1 INTRODUCTION

Decades of research in the field of automatic speech recognition (ASR) brought numerous methods to improve the recognition performance by increasing the robustness against variability of speech signals. Several of these methods are inspired by the principles of human speech perception, which is motivated by the fact that the robustness of human recognition performance exceeds by far the robustness of ASR performance even in acoustically optimal conditions (Lippmann, 1997; Cooke and Scharenborg, 2008; Meyer et al., 2011b). The sources of variability in spoken language can be categorized into extrinsic sources (e.g., background noise, the room acoustics, or distortions of the communication channel) and intrinsic sources, which are associated with the speech signal itself (e.g., the talkers' speaking style, gender, age, mood, etc.). Compared to the human auditory system, ASR was found to be far less robust against both types of variability (Lippmann, 1997; Benzeghiba et al., 2007).

In this study, the focus lies on the improvement of feature extraction by using a set of physiologically inspired filters (Gabor filters), which is applied to a spectro-temporal representation of the speech signal. In order to choose a set of filters suitable for ASR tasks, a filter bank is defined and used to extract a wide range of spectro-temporal modulation frequencies (MF) from the signal, while at the same time limiting the redundancy on feature level.

Most state-of-the-art ASR systems perform an analysis of short-time segments of speech and use spectral slices, typically calculated from 25 ms segments of the signal as feature input. The most successful implementations of such spectral processing are Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) and perceptual linear prediction features (PLPs) (Hermansky, 1990). These features are usually concatenated with their first and second order discrete temporal derivation (delta and double-delta features) to incorporate information about the temporal dynamics of the underlying signal on feature level. The PLP feature extraction was later refined by performing RASTA (RelAtive SpecTrA) processing, which effectively suppresses temporal fluctuations that correspond to background noise or changes of the transmission channel (Hermansky and Morgan, 1994). The idea of using temporal cues was implemented in form of temporal pattern (or TRAPS) features, which were found to increase robustness of ASR systems in

noisy environments (Hermansky and Sharma, 1999). These approaches suggest that both spectral and temporal integration of a spectro-temporal representation of the signal may be useful for speech processing, which has therefore motivated studies that incorporate such spectro-temporal processing for ASR.

From a physiological point of view, it seems worthwhile to feed spectro-temporal features to ASR engines, since several studies indicate that a similar processing is performed by the auditory system: These findings indicate that some neurons in the primary auditory cortex of mammals are explicitly tuned to spectro-temporal patterns. For example, Qiu et al. (2003) used specific spectro-temporal patterns to identify spectro-temporal receptive fields (STRFs) in the auditory cortex in cats. An STRF is associated with a particular neuron or a group of neurons; it is an estimate for the spectro-temporal representation of the sound stimulus that optimally "drives" the neuron. More recent findings show that spectro-temporal representations of human speech found in the primary auditory cortex of ferrets are well-suited to distinguish phonemes (Mesgarani et al., 2008). The observation that such information is encoded in auditory processing stages serves as motivation for the explicit use of this type of representation in speech pattern recognition.

Different types of spectro-temporal features for ASR have been investigated in the past. Ezzat et al. (2007a) and Bouvrie et al. (2008) analyzed spectro-temporal patches with a 2D discrete cosine transform. They used this representation as a tool for speech analysis and for the extraction of robust features. Heckmann et al. (2008) and Domont et al. (2008) employed spectro-temporal patches to derive STRFs from artificial neurons. Another type of spectro-temporal features originates directly from the modeling of the patterns observed in the STRFs in the auditory cortex in cats.

Qiu et al. (2003) modeled these patterns with two-dimensional Gabor functions. This motivated Kleinschmidt and Gelbart (2002a) to apply Gabor filters to the problem of ASR, with the aim of explicitly incorporating spectro-temporal cues on the feature level. An example of a two-dimensional Gabor filter is shown in Fig. 2.1. These filters were also shown to be suitable for the analysis of speech properties [e.g., for the distinction of plosives, fricatives and nasals (Ezzat et al., 2007b)]. Mesgarani et al. (2006) found that the use of auditory Gabor features improves classification results for speech/nonspeech detection in noisy environ-

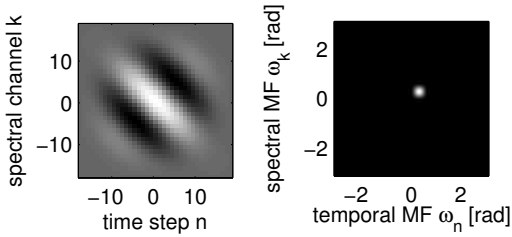

Figure 2.1: *2D Gabor filter. (left) Real part. Black and white shading correspond to negative and positive values, respectively. (right) Absolute values of the filter's transfer function in the modulation domain. White shading corresponds to high amplitude.*

ments. The extraction of features requires a set of Gabor filters in order to capture information about spectral, temporal and spectro-temporal patterns.

One of the challenges when applying Gabor filters to speech-related tasks is finding a suitable set of filters from the vast number of parameter combinations and which extracts relevant information from the spectro-temporal representation. Standard back-ends such as Hidden Markov Models (HMMs) using Gaussian Mixture Models (GMMs) often require the components of input features to be decorrelated, and computational restrictions make the use of very large vectors (with more than 1000 components) difficult.

In the past, different methods were proposed to cope with this challenge. Kleinschmidt (2003) and Meyer and Kollmeier (2011a) used a stochastic feature selection algorithm [the Feature Finding Neural Network (FFNN) (Gramss, 1991)] that was initialized with a random set of 80 filters. Based on the performance on a simple recognition task (i.e., isolated digits), filters that were found to decrease ASR performance were discarded and replaced with a new random filter, which eventually resulted in a set that was found to increase the noise robustness for the recognition of noisy digit strings. Improvements over the MFCC baseline were obtained by using Gabor features as input to a Tandem system that consists of an artificial neural net (or multi-layer perceptron, MLP). The MLP transformed the Gabor input features into posterior probabilities for

phonemes. These posteriors were then decorrelated and used as input to a conventional GMM/HMM classifier.

A different approach is to consider the outputs of the different Gabor filters as feature streams, and start with a very high number of filters (up to tens of thousands compared to the 80 filters mentioned before), and subsequently merging filter outputs that are organized in streams with neural nets. A merger MLP was used to combine isolated streams, and a PCA was applied to its output. This approach was used by Chi et al. (2005), Zhao and Morgan (2008), and Mesgarani et al. (2010).

These studies have shown that spectro-temporal information helps to increase the robustness of ASR systems. Meyer and Kollmeier (2011a) assumed that the benefits observed for spectro-temporal features (compared to purely spectral feature extraction) arise from a local increase of the SNR since the Gabor functions serve as matched filters for specific spectro-temporal structures in speech, such as formant transitions. However, for several studies (Kleinschmidt and Gelbart, 2002a; Meyer and Kollmeier, 2011a), a different database was used for MLP training than for the task for which results were reported, and it is unclear if this additional training material might result in an advantage over setups that do not make use of additional training data. Since all of these studies use the combination of MLPs and PCA, the physical meaning (in terms modulation frequencies) is not directly interpretable from the features that are ultimately fed to the back end. However, when using front-ends as a tool for analysis that might give a hint on what kind of input data is actually helpful, the physical interpretability is a desirable feature.

The aims of this study are to design a filter bank of spectro-temporal filters that are applicable to extract ASR features, and to use these for an analysis of parameters relevant for speech recognition based on spectro-temporal features. Among the design decisions are the number of filters considered for the filter bank, their phase sensitivity, and the spectral and temporal modulation frequencies to be used. Such a 2D filter bank can then be employed to analyze the relative importance of modulation frequencies. Kanedera et al. (1999) performed a series of experiments that quantified the importance of purely temporal modulation frequencies for ASR. One of the results is that temporal modulations in the range of 2 Hz to 16 Hz play the dominant role for ASR performance. In this study, this analysis is extended to spectral and spectro-temporal modulation frequencies by performing ASR experiments when specific modulation

frequencies are disregarded. Nemala and Elhilali (2010) analyzed the contribution of different temporal and spectral modulation frequencies for robust speech/non-speech classification and found temporal modulations from 12 Hz to 22 Hz and spectral modulations from 1.5 to 4 cycles/octave to be particularly useful to achieve robustness in highly noisy and reverberant environments.

We then evaluate the robustness of these features in the presence of intrinsic and extrinsic sources of variability, and compare them to a range of spectral feature types that are commonly applied in ASR. ASR performance in the presence of additive noise and varying channel characteristics is investigated with two experimental setups (i.e., the widely used Aurora2 digit recognition task that employs the HTK back end, and the Numbers95 task for which a state-of-the-art backend was used). The effect of intrinsic variation is explored using a phoneme detection task (in which phonemes are embedded in short nonsense utterances).

The structure of this paper is reflected by these aims: We first present the design decisions for the Gabor filter bank (Sec. 2.2), how it is applied to feature extraction, and which modulation frequencies were found to be relevant for this ASR task (Sec. 2.2.1). Section 2.2.3 presents the corresponding results. The experiments that investigate the sensitivity of spectro-temporal and baseline features against extrinsic and intrinsic variability are presented in Sec. 2.3.2. Sections 2.3.3 and 2.4 present the results, the discussion and conclusions.

2.2 GABOR FILTER BANK FEATURES

This section describes the design of the Gabor filter bank, the choice of its parameters, and the calculation of the Gabor filter bank features (GBFB). With these features, we perform an analysis of the importance of phase information in spectro-temporal pre-processing, evaluate the effect of selecting specific modulation frequencies.

2.2.1 Calculation of the GBFB features

An overview of the feature extraction scheme with the Gabor filter bank process is illustrated in Fig. 2.2. First, a Mel-spectrogram is calculated from the speech signal using an implementation of the ETSI Distributed

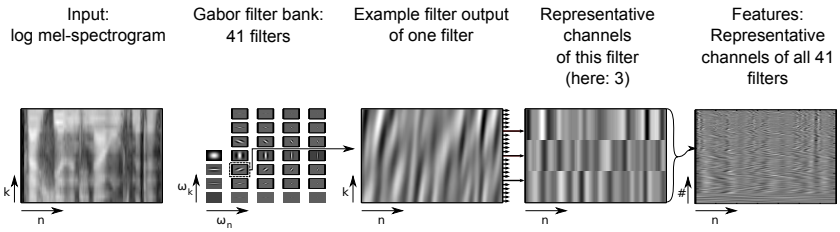

Figure 2.2: *Illustration of the Gabor filter bank feature extraction. The input log Mel-spectrogram is filtered with each of the* 41 *filters of the Gabor filter bank. An example filter output is shown. The representative channels of this filter output are selected and concatenated with the representative channels of the other* 40 *Gabor filters. The resulting* 311*-dimensional output is used as feature vector.*

Speech Recognition Standard (ETSI, 2003, Standard 201 108 v1.1.3). This standard defines the calculation of a Mel-spectrogram that consists of 23 frequency channels with center frequencies in the range from 124 Hz to 3657 Hz. The calculation is based on frames of 25 ms length, while the temporal resolution is 100 frames/s. The spectrogram incorporates a Mel-frequency scale that is logarithmic for frequencies above 1 kHz and therefore mimics the mapping of frequencies to specific regions of the basilar membrane in the inner ear. Since the frequency mapping is not strictly logarithmic (with approximately linear at frequencies below 800 Hz), the spectral modulation frequencies are specified in cycles per channel. The absolute output values of the spectrogram are compressed with the logarithm, roughly resembling the amplitude compression performed by the auditory system. The spectrogram is then processed with the filters from the GBFB, which are introduced in Sec. 2.2.1.1, by calculating the two-dimensional convolution of the spectrogram and the filter. This results in a time-frequency representation that contains patterns matching the modulation frequencies associated with a specific filter. The filtering process is illustrated in Fig. 2.3, which shows the original spectrogram, a sample filter, and the filter output.

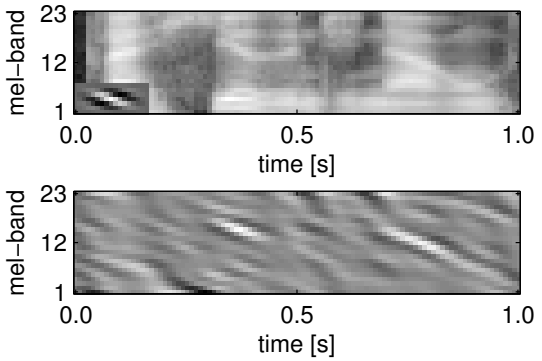

Figure 2.3: *Illustration of the filtering process with a Gabor filter. (top) Mel-spectrogram of the German sentence "Gleich hier sind die Nahrungsmittel" (The food is right over here) that exhibits spectro-temporal (diagonal) structures that arise from vowel transitions and Gabor filter (real part shown in the lower left corner of the spectrogram). (bottom) 2D filter output obtained by calculating the convolution of the Mel-spectrogram and the real part of the filter. White shading corresponds to high energy on the logarithmically scaled color encoding.*

2.2.1.1 Gabor filter bank

The localized complex Gabor filters are defined in Eq. (2.1), with the channel and time-frame variables k and n; k_0 denoting the central frequency channel; n_0 the central time frame; ω_k the spectral modulation frequency; ω_n the temporal modulation frequency; ν_k and ν_n the number of semi-cycles under the envelope in spectral and temporal dimension; and ϕ an additional global phase:

$$h_b(x) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi x}{b}\right) & -\frac{b}{2} < x < \frac{b}{2} \\ 0 & \text{else} \end{cases}, \tag{2.1a}$$

$$s_\omega(x) = \mathrm{e}^{\mathrm{i}\omega x}, \tag{2.1b}$$

$$g(k_0, n_0, \omega_k, \omega_n, k, n, \nu_k, \nu_n, \phi) = \underbrace{s_{\omega_k}(k-k_0)\, s_{\omega_n}(n-n_0)}_{\text{carrier function}} \cdot \underbrace{h_{\frac{\nu_k}{2\omega_k}}(k-k_0)\, h_{\frac{\nu_n}{2\omega_n}}(n-n_0)}_{\text{envelope function}} \cdot \underbrace{\mathrm{e}^{\mathrm{i}\phi}}_{\text{phase}} . \tag{2.1c}$$

A Gabor filter is defined as the product of a complex sinusoid carrier [Eq. (2.1b)] with the corresponding modulation frequencies ω_k and ω_n, and an envelope function [Eq. (2.1a)]. For purely temporal and purely spectral modulation filters ($\omega_n = 0$ or $\omega_k = 0$) this definition results in filter functions with infinite support. For that reason the filter size of all filters is limited to 69 channels and 40 time frames. These limits correspond roughly to the maximum size of the spectro-temporal filters in the respective dimensions. Due to the linear relation between the modulation frequency and the extension of the envelope, all filters with identical values for ν_k and ν_n are constant-Q filters.

Since relative energy fluctuations are of special interest for the classification of speech, the DC bias of each filter is removed. This is achieved by subtracting a normalized version of the filter's envelope function from the filter function, so that their DC values cancel each other out. Filters that are centered near the edges of the spectrogram usually do not lie completely within the boundaries of the spectrogram. Hence, the DC removal is applied for all center frequencies separately to avoid artifacts. The effect of the DC removal is that the resulting representation is independent of the global signal energy. Since a removal of the mean on a logarithmic energy scale is the same as dividing by it on a linear scale, this corresponds to a normalization. While cepstral coefficients normalize spectrally, and RASTA processing and discrete derivatives normalize temporally, DC-free Gabor filters naturally normalize in both directions.

The filter bank is designed with the aim of evenly covering the modulation frequencies in the modulation transfer space as schematically illustrated in Fig. 2.4. Cross-sections of the filter transfer functions along the x axis and y axis of this representation are depicted in Fig. 2.5.

The distribution of spectro-temporal modulation frequencies is defined by Eq. (2.2), which ensures that adjacent filters exhibit a constant overlap in the modulation transfer domain:

$$\omega_{\mathrm{x}}^{i+1} = \omega_{\mathrm{x}}^{i} \frac{1 + \frac{c}{2}}{1 - \frac{c}{2}}, \tag{2.2a}$$

$$c = d_{\mathrm{x}} \frac{8}{\nu_{\mathrm{x}}}. \tag{2.2b}$$

The advantage of this definition is that each filter accounts for a different combination of spectral and temporal modulation frequencies (ω_n, ω_k) and thus has limited correlation with the other filters.

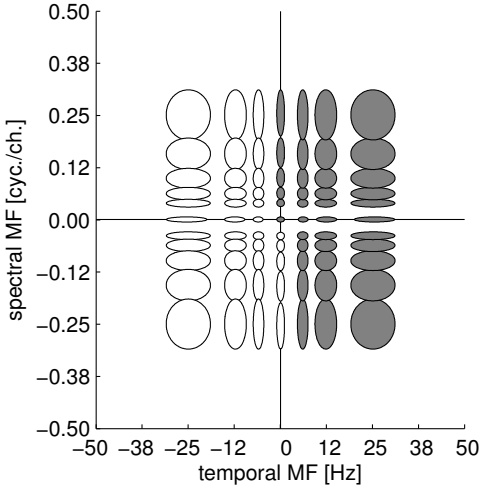

Figure 2.4: *Illustration of the distribution and size of the transfer functions of the Gabor filter bank filters. Each circle/ellipse corresponds to one Gabor filter and is centered on its center frequency. The circles/ellipses mark the -1 dB level of the filters. With the exception of filters on the axis, the relation between the center modulation frequency and the bandwidth of its pass-band is proportional. Since only the real part of the filter output is considered for feature extraction, centrally symmetric filters yield identical outputs. Therefore, only the filters that correspond to the filled circles/ellipses are used for feature extraction.*

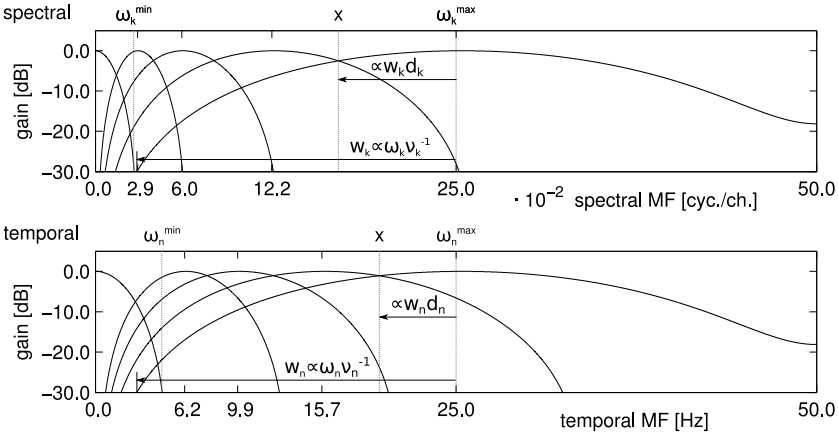

Figure 2.5: *Cross-section along the spectral and the temporal axis of the modulation transfer space showing the gains of the individual transfer functions. The width of a filter w is proportional to the center modulation frequency ω and anti-proportional to the number of half-waves under the envelope ν, and is indicated here for the highest modulation frequency. The distance to the point where two adjacent filters have equal gains (marked for the filter with the highest modulation frequency with an x) is proportional to the width and the distance factor d. Note that the distance parameter d also controls the overlap between adjacent filters. In the upper panel d_k is chosen* 0.3*, where in the lower panel d_n is* 0.2*.*

Figure 2.5 also explains the meaning of the parameters of the GBFB. The upper and lower bounds for the modulation frequencies are given by $\omega^{\max}$ and $\omega^{\min}$. The width of a filter w is proportional to the center modulation frequency ω and anti-proportional to ν, which results in constant-Q filters. The distance to the point where two adjacent filters have equal gains (marked with an x) is proportional to the width and the distance factor d. This factor is used to adjust the overlap of adjacent filters, with small values for d resulting in a large overlap and with $d = 1$ corresponding to a coincidence of the first zeros of adjacent filters. The redundancy of the filter outputs due to their overlap can thus be controlled by the distance parameter d.

The modulation frequencies (ω_n, ω_k) can assume positive or negative values. The signs determine the spectro-temporal direction the filter is tuned to. Filters with only one negative modulation frequency correspond to rising spectro-temporal patterns, while other filters correspond to falling spectro-temporal patterns. Since the feature extraction uses the real part of the filter outputs, only filters with positive modulation frequencies and their symmetric versions with one sign inverted are considered, as inverting both signs would yield identical filters. This relation is illustrated in Fig. 2.4. Only the filters that correspond to the filled circles/ellises are used. The corresponding filters are depicted in Fig. 2.6.

2.2.1.2 Selection of representative frequency channels

When using the filter output of all 41 filters, the resulting feature vector is relatively high-dimensional with 23 (frequency channels) $\times$ 41 (filters). We reduce the number of feature components by exploiting the fact that the filter output between adjacent channels is highly correlated when the filter has a large spectral extent (cf. Fig. 2.2). Since highly correlated feature components can result in reduced ASR performance (especially when only a small amount of training data is available), a number of representative channels is selected by subsampling the 23-dimensional filter output for each filter. The central channel, corresponding to about 1 kHz, is selected for all feature vectors because the most important cues for ASR are more likely to be found in the center rather than at the edges of the spectrum. Additionally, channels with an approximate distance of a multiple of 1/4 of the filter width to the center channel are included. The value 1/4 is motivated by the sampling theorem in the same way as

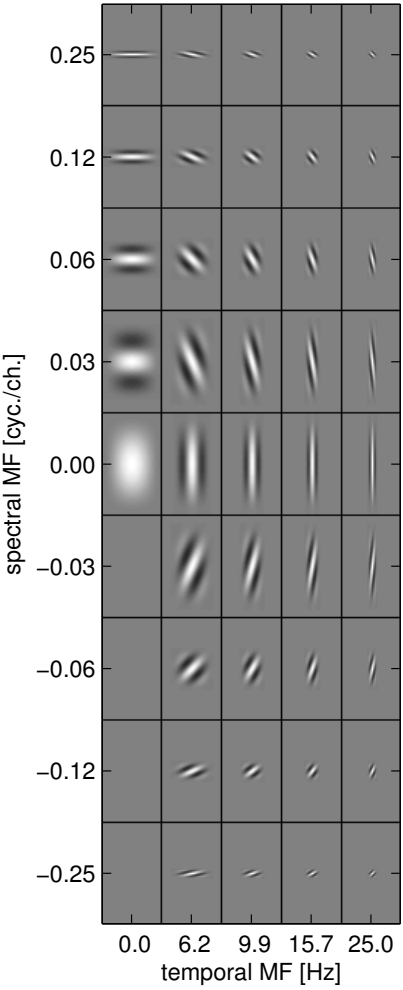

Figure 2.6: *Real part of the* 41 *Gabor filters used for the Gabor filter bank feature extraction in time-frequency domain. Black and white shading corresponds to negative and positive values, respectively.*

the minimum window overlap that is needed in a spectrogram for perfect reconstruction.

For filters with the lowest spectral extent, all 23 components are selected for the feature vector, while for the largest filters only a single component (the central frequency channel) is kept. An example with three selected channels is shown in Fig. 2.2. This selection scheme reduces the filter bank output to 311 dimensions, which is referred to as GBFB features. Alternatively, a principal component analysis (PCA) may be applied to the full filter bank output, which has the same effect as the channel selection (i.e., the decorrelation of feature components, and the reduction of dimensionality). We therefore test the application of PCA to the filter bank output and compare the results to the proposed scheme of channel selection.

2.2.1.3 Implementation

The calculation of GBFB features results in higher computational load compared to standard front-ends (by a factor of 80 compared to MFCC features), which may be an issue on small-footprint systems. However, GBFB feature calculation can be performed in real-time on a single-core standard PC, and with the current development of dual- and many-core processors, considerable speedups can be achieved by parallelizing the 2D convolutions of the filters. A reference implementation of the GBFB feature extraction in MATLAB is available online[1].

2.2.2 Experiments

Before describing the experiments with the Gabor filter bank, the Aurora 2 framework, the automatic speech recognition framework that is used in all of the following experiments in this section to determine performance and robustness, is introduced.

2.2.2.1 Aurora 2: Digits in noise recognition task

To evaluate robustness against extrinsic variability the Aurora 2 framework is used (Pearce and Hirsch, 2000). It consists of the Aurora 2 speech database, a reference feature extraction algorithm (MFCC), a recognizer setup [Hidden Markov Toolkit (HTK) (Young et al., 2001)],

[1] URL: http://medi.uni-oldenburg.de/GBFB

and rules for training and testing. The recognition task is the classification of connected digit strings with artificially added noise. The database contains digits spoken by native English speakers and everyday noise signals recorded at 8 different sites (subway, babble, car, exhibition, restaurant, street, airport, trainstation). The test set consists of digits with noises added at different SNRs ranging from 20 dB to −5 dB. The standard features used in the Aurora 2 framework are MFCC features with their first and second discrete derivative. For speech data modeling the HTK recognizer employs Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs).

In the Aurora 2 framework, two training and three test conditions are defined: Clean training uses only clean utterances, while for multi-condition training a mixture of noisy (subway, babble, car, exhibition) and clean digit strings is used. Test set A contains noises also used for training, while for test set B unknown noise types (restaurant, street, airport, station) are used. Test set C contains samples that have been filtered with a different transfer function than the samples of test set A, B, and the training data to simulate a change in communication channel properties.

The HMM back end is configured according to the Aurora 2 guidelines for all feature types: The number of HMM states per word is 18, the number of Gaussian mixtures per state is three; an additional tuning of the back end is not performed. Although tuning might improve results especially when the feature dimension strongly differs from the dimensionality of baseline features, we keep the parameters for reasons of comparability with other studies that use the Aurora 2 framework. For all features the number of time frames is kept constant, because skipping a few frames at the beginning and the end of the utterances improves the performance as it narrows the region to where speech occurs.

The experiments are carried out with different feature types to compare their robustness with respect to the effect of the mismatches between the training and the test data, represented by test sets A, B, and C. The results obtained with the Aurora 2 setup consist of the results for multi-condition and clean training. Word recognition accuracies (WRA) in percent are calculated for each noise condition and for each signal-to-noise ratio (SNR) separately. We also present the relative reduction of the word error rate (WER), which is calculated by determining the relative reduction of error $\mathrm{WER} = 1 - \mathrm{WRA}$ for each SNR/noise condition (with

Table 2.1: *GBFB parameter values used for feature extraction in comparison with values derived from parameters of the baseline features. Spectral modulation frequecies ω_k are reported in cycles per channel, and temporal modulation frequencies ω_n are reported in Hz.*

	Parameters (or their approximated analogues)							
Features	$\omega_k^{\min}$	$\omega_k^{\max}$	$\omega_n^{\min}$	$\omega_n^{\max}$	ν_k	ν_n	d_k	d_n
GBFB	0.0254	0.2500	4.38	25	3.5	3.5	0.3	0.2
MFCC WI007	≈ 0.022	≈ 0.28	0.0	50	$1-13$	$\approx 1-3$	≈ 0.13	–
RASTA-PLP	–	–	≈ 2.6	≈ 20	–	–	–	–

SNRs ranging from 0 to 20 dB) and averaging over those improvements. The average relative improvements of each noise condition and of test set A, B, and C are calculated to differentiate the effect of different types of mismatches between training and test data. Furthermore, the average word recognition accuracy and the relative improvement for each SNR is calculated.

2.2.2.2 GBFB parameters

This section describes how several of the parameters of the filter bank were chosen. We also compare this choice to the corresponding parameters of the baseline features. Given the structure of the filter bank (that defines, for example, the position of filters in temporal and spectral dimension given a spacing between those filters), we are left with eight parameters that need to be specified: The lowest and highest temporal and spectral modulation frequencies $(\omega_n^{\max}, \omega_k^{\max}, \omega_n^{\min}, \omega_k^{\min})$, the number of periods used for the filters (ν_n, ν_k), and the overlap of adjacent filters (d_k, d_n). The initial values for these parameters are chosen based on the corresponding values of the baseline features (cf. Table 2.1). For instance, the spectral modulation frequencies associated with the baseline MFCCs range from 0.022 to 0.28 cycles/channel, and the parameters for GBFB features were chosen accordingly. The same is true for the temporal modulation frequencies that are relevant in RASTA-processing of signals. Further optimization was carried out by performing a series of ASR experiments varying the parameters one after another on the Aurora 2 task, finding the parameters that result in best overall performance. The optimization was carried out with a fixed phase setting of $(\phi = 0)$, because this way

the maximum amplitude of the filters coincides with the center of the filter independently of its modulation frequency. The parameters were not optimized in any particular order which could have led to finding a local optimum in the parameter space.

To test if GBFB features are overfitted to the Aurora 2 task by the selection of a specific set of parameters, variations of all parameters to the best performing set are evaluated. From the default set of parameters in Table 2.1, each parameter is set to different values, covering a wide range of the plausible parameter space. The selection of frequency channels (Sec. 2.2.1.2) was not optimized. Instead we apply the outlined scheme to the full output of the filter bank, and compare the results to transforming the full output with a PCA. The results are presented and discussed in Sec. 2.2.3.1.

2.2.2.3 Importance of GBFB phase information

From the output of the Gabor filter bank, either the real or imaginary part, or the absolute values may be used. Using the imaginary part of the output is equivalent to choosing the parameter $\phi = \pi/2$, and effectively using the filters as edge detectors of spectro-temporal events. The absolute values of the output are less sensitive to the exact spectro-temporal location. The phase of the Gabor filters does not matter in this case. To test the importance of the phase information of the filter bank output for robust ASR the performance of the real part, the imaginary part and absolute values of the filter output is compared on the Aurora 2 task. The results are presented and discussed in Sec. 2.2.3.2.

2.2.2.4 Relative importance of specific modulation frequencies

In order to evaluate the importance of specific modulation frequencies for ASR, a band-stop experiment is performed that quantifies the contribution of specific combinations of spectral and temporal modulation frequencies (ω_k, ω_n) to the overall ASR performance. For this evaluation, the feature components associated with a specific modulation frequency are removed from the output of the Gabor filter bank. This approach results in 41 different reduced filter sets. Since the number of center frequencies associated with a specific spectro-temporal modulation frequency varies (cf. Sec. 2.2.1.2), the number of dimensions removed from the GBFB output ranges from 0 to 22. When the accuracy decreases when omitting

filters with a particular modulation frequency, these filters are likely to extract relevant information that is not covered by the remaining Gabor filters. On the other hand, if the accuracy increases when filters are omitted, this indicates that the filters capture information that is either covered by the remaining filters or not relevant for this specific speech recognition task.

The importance of the filters is evaluated with the Aurora 2 task, since this speech material is expected to exhibit a more natural distribution of temporal modulation frequencies compared to the very short utterances from the OLLO database. The Aurora 2 recognizer is trained and tested with each reduced feature representation. The results are presented and discussed in Sec. 2.2.3.3.

2.2.3 Results and discussion

2.2.3.1 GBFB parameters

Overall recognition performance in % WRA and % relative reduction of the WER over the MFCC WI007 baseline for variations of the GBFB parameters of Table 2.1 are presented in Table 2.2. The recognition performance for the GBFB features with altered parameters changes compared to the original set. For some parameters the best values are different for clean and multi-condition training. Hence, the set of parameters that is used for feature extraction is a trade-off between performance for clean training and performance with multicondition training and each could be improved further by selecting different parameters. With many of the changes to the original parameter set, the GBFB features still improve the MFCC WI007 baseline. Some parameters affect more the overall performance, some affect more the relative improvement over the baseline, but there is no clear trend.

The presented results were obtained by selecting frequency channels from the filter output as described in Sec. 2.2.1.2. In order to evaluate if a decorrelation and dimension reduction with a PCA should be preferred over channel selection, we apply a PCA either to the full filter bank output (i.e., channel selection is not performed) or the 311-dimensional GBFB features. In each case, the transformation statistics is obtained from the corresponding (clean or multi-condition) training material, and the feature dimension is reduced to 39 (the dimension of the baseline features). From each dimension of the data the mean is removed and

Table 2.2: *Overall word recognition accuracies (WRA) and relative reduction of word error rates (relative improvement) compared to the MFCC baseline with clean and multi condition training on the Aurora 2 task for various modifications to the GBFB parameters.*

Parameter		ν_k				ν_n			
Values		2.5	3.0	4.0	4.5	2.5	3.0	4.0	4.5
WRA [%]	clean	56.7	63.0	68.2	70.0	69.6	68.6	62.6	61.2
	multi	86.4	88.1	86.5	87.3	83.3	86.7	87.6	87.0
Rel. Imp. [%]	clean	−18.8	14.5	33.9	37.7	30.8	33.0	14.3	10.0
	multi	−2.8	18.6	9.2	5.27	−48.0	−2.2	13.4	10.2
Parameter		d_k				d_n			
Values		0.1	0.2	0.4		0.1	0.3	0.4	
WRA [%]	clean	63.6	65.1	67.4		65.3	67.0	64.8	
	multi	87.1	87.8	87.9		87.8	87.5	84.6	
Rel. Imp. [%]	clean	22.7	27.0	28.9		28.1	29.4	11.4	
	multi	12.0	14.0	12.5		17.5	11.4	−17.7	
Parameter		$\omega_k^{\max}$ $10^{-2}[\frac{\text{cyc}}{\text{ch}}]$				$\omega_n^{\max}$ [Hz]			
Values		18.75		12.5		18.75		12.5	
WRA [%]	clean	62.1		61.9		69.3		69.0	
	multi	86.8		85.0		87.5		88.2	
Rel. Imp. [%]	clean	16.0		14.2		34.2		33.4	
	multi	3.9		−5.0		7.9		7.4	
Parameter		$\omega_k^{\min}$ $10^{-2}[\frac{\text{cyc}}{\text{ch}}]$				$\omega_n^{\min}$ [Hz]			
Values		1.9	3.8	7.61		2.19	3.5	8.75	
WRA [%]	clean	66.2	61.4	59.0		67.8	65.5	65.7	
	multi	88.1	86.5	89.3		88.9	88.7	87.6	
Rel. Imp. [%]	clean	28.4	12.0	−8.3		35.0	28.2	26.3	
	multi	16.6	6.1	24.1		10.5	17.2	15.5	

the variance is normalized before calculating the PCA coefficients. The results are shown in Table 2.3.

The application of a PCA to the *full* filter bank output results in recognition rates below the GBFB features and the MFCC baseline. When a PCA is applied to the GBFB features with representative channels, the absolute score for clean training is improved, whereas multi-condition results are better with GBFB features. The relative improvements over the baseline are slightly higher with the original GBFB approach.

Table 2.3: *Comparison of GBFB features that incorporate the selection of frequency channels from the filter output (GBFB), processing the full filter bank output with a PCA ($^{\text{GBFB}}_{\text{full}}$ and PCA) and application of a PCA to the GBFB features (GBFB and PCA). The recognition performance is presented in word recognition accuracies in % and as relative improvement over the MFCC baseline for clean (c) and multi (m) condition training.*

Method		GBFB	GBFB and PCA		$^{\text{GBFB}}_{\text{full}}$ and PCA	
PCA Traindata		-	clean	multi	clean	multi
WRA [%]	c	66.2	**69.6**	64.5	56.5	45.4
	m	**88.1**	82.9	84.6	85.0	86.2
Rel.Imp. [%]	c	**28.4**	28.3	−23.6	−28.6	−79.0
	m	**16.2**	−47.3	14.2	−14.1	−9.2

We therefore argue that the direct use of GBFB features should be preferred over PCA-transformed features, since GBFB features are easier to calculate, produce slightly better results on average, and the physical meaning of feature components is retained (i.e., each feature component is associated with a modulation frequency, which enables experiments such as the evaluation of the contribution of such physical parameters to ASR).

The results of the parameter variation and the PCA show that the feature extraction can be optimized for a specific condition. For multi-condition training for example, even less robust patterns may serve for the recognition, as their uncertainty is known. These patterns could be matched by Gabor filters with diverse shapes. In this sense, the GBFB structure limits the fitting to a specific task by greatly reducing the degree of freedom of the feature extraction in contrast to a set of independent Gabor filters. The GBFB features project the log Mel-spectrogram to a over-complete basis of a subspace of the log Mel-spectrogram. The sub-space is limited by the lower and upper bounds for the modulation frequencies ($\omega_k^{\min}$ to $\omega_k^{\max}$ and $\omega_n^{\min}$ to $\omega_n^{\max}$). Its degree of over-completeness is adjusted by the distance parameter d and the shape of the basis functions is determined by ν.

It is likely that for different tasks different sets of parameters are optimal, as it is also the case with traditional features. However, we found that none of the parameters of this very generic projection is critical to outperform the MFCC baseline. Nonetheless, ASR systems are

Table 2.4: *Average word recognition accuracies (WRA) and relative reduction of word error rates (relative improvement) compared to the MFCC baseline with clean (c) and multi (m) condition training on the Aurora 2 task for the real part, the imaginary part and the absolute values of the GBFB features.*

Modification		None (real)	Imaginary	Absolute
WRA [%]	c	66.2	**67.0**	48.8
	m	**88.1**	87.8	81.8
Rel. Imp. [%]	c	28.4	**31.6**	−59.7
	m	**16.2**	10.7	−42.2

non-linear and complex so that the front end and the back end cannot be judged independently. Back ends make strong assumptions about the feature's statistical characteristics, which lead to degraded recognition performance if ignored. A remaining question is if the improvements made with GBFB features for the Aurora 2 task will translate to other ASR setups. For that reason the GBFB features that are adapted to work well with the GMM/HTK back end of the Aurora 2 task are evaluated on another recognition task with a different back end in Sec. 2.3.

2.2.3.2 Importance of GBFB phase information

Overall recognition performance in % WRA and % relative reduction of the WER over the MFCC WI007 baseline for the real part, the imaginary part and the absolute values of the GBFB features are presented in Table 2.4. The accuracies obtained with the real and imaginary part are in the same range, whereas the performance with absolute values (for which the location of spectro-temporal events is smeared out) is reduced considerably. This indicates that phase information is an important factor for ASR, and should be considered in spectro-temporal feature extraction. Since the real-valued filter output performs slightly better than features based on imaginary filters on average, we use the real output for ASR experiments.

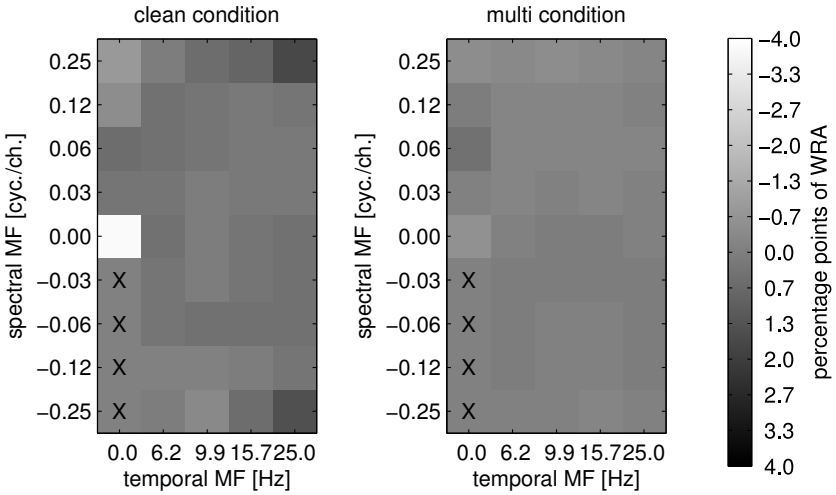

Figure 2.7: *Differences in overall accuracy on the Aurora 2 digit recognition task when omitting the output of filters with a particular spectro-temporal modulation frequency for multi and clean condition training. The difference in accuracy is encoded in grayscale and displayed at the position of the corresponding center modulation frequency of the omitted filter. Filters that are not used for feature extraction are marked with an X.*

2.2.3.3 Relative importance of specific modulation frequencies

In this section, the digit recognition performance is determined based on reduced filter sets, for which a spectro-temporal modulation frequency is omitted as described in Sec. 2.2.2.4. The aim of this experiment is to estimate the relative importance of specific modulation frequencies. Figure 2.7 shows the difference between the recognition scores obtained with the original and the reduced features. Therefore, low values correspond to filters with a relatively high contribution to the recognition scores.

The patterns observed in Fig. 2.7 show a symmetry with respect to upward and downward filters (i.e., those with negative and positive modulation frequencies, respectively). On average, filters tuned to upward spectro-temporal patterns and filters tuned to downward spectro-temporal

patterns appear to be equally important for the recognition with clean training. For multi-condition training the effect of omitting a filter is smaller (Fig. 2.7, right), but symmetry of upward and downward filters is not affected. The most important feature is the output of the DC filter, which encodes the level of the recording averaged over about 300 ms. The DC feature may be seen as a simple voice activity detector, and its information is not encoded in any of the other feature channels, as these do not have a DC component. Its exclusion reduces the average word recognition accuracy by about 4 percentage points, with multi-condition training it causes a drop by approximately 0.6 percentage points. The most important modulation frequency belongs to the purely spectral filter ($\omega_n = 0\,\text{Hz}$) with the highest modulation frequency ($\omega_k = 0.25$ cycles/channel). It accounts for the finer spectral structure of the log Mel-spectrogram. We assume that this is the filter that best extracts information about voicing, as voicing features are represented by localized patterns that usually do not exceed two frequency channel and do not exhibit strong temporal changes.

Several filters have a detrimental effect on the overall performance, since their removal from the feature vector results in an *increase* of recognition performance: Omitting the filters with the highest modulation frequencies ($\omega_k = \pm 0.25$ cycles/channel and $\omega_n = 25\,\text{Hz}$) improves the recognition performance by about 2 percentage points with clean training. The spectral filter ($\omega_n = 0\,\text{Hz}$) with the lowest modulation frequency ($\omega_k = 0.03$ cycles/channel) also has a detrimental effect. This filter accounts for the very coarse spectral shape of the log Mel-spectrogram averaged over about 300 ms. It extracts mainly information about the spectral color of the communication channel. The improvement in overall scores upon deletion of specific components indicates that feature selection may further improve the recognition accuracy.

Kanedera et al. (1999) found that temporal modulation frequencies below 2 Hz and above 16 Hz may be detrimental for specific ASR tasks. The temporal modulation center frequencies used for the filter bank range from 6.2 Hz to 25 Hz and are subdivided into spectro-temporal upward and spectro-temporal downward filters. With GBFB features, an upper limit of about 18 Hz (cf. Table 2.2) seems to improve performance with clean condition training from 66.2% to 69.3% overall but reduce performance with multi-condition training from 88.1% to 87.5%. The range of modulation frequencies used with GBFB features is higher than

the range found by Kanedera et al. (1999). Some temporal modulation frequencies are only beneficial in combination with certain spectral modulation frequencies. Nemala and Elhilali (2010) found temporal modulation frequencies from 12 Hz up to 22 Hz to be useful for robust speech/non-speech recognition in an experiment that considered spectral and temporal modulation frequencies. The range of modulation frequencies used with the GBFB is in line with these findings. It is possible that an interaction between spectral and temporal modulation frequencies results in a shift of the specific frequencies important for ASR.

The most frequent temporal modulation frequency in speech is 4 Hz, but it was not found to be of particular importance for the recognition of connected digits that spectro-temporal filters tuned to 4 Hz existed at the feature level. This does not mean that it is of no importance at all, since temporal modulation frequencies below 6.2 Hz are captured by the purely spectral filters and the back-end models changes of this rate. An example for such a filter is the DC filter that changes with a temporal rate of up to about 4 Hz (cf. filter transfer function in Fig. 2.5) and plays an important role.

Another factor that might affect the overall recognition accuracy is the number of individual feature components associated with a spectro-temporal modulation frequency: The results of the filtering process is a spectro-temporal output with 23 frequency channels; in most cases, not all of these channels are included in the feature vector to avoid a high redundancy of feature components. The number of selected channels ranges from 1 (for low values of ω_k) to 23 (for high values of ω_k). Since modulation filters are disregarded in the band-stop experiment, the number of components ranges from 288 to 310, which might have an effect on the overall performance.

2.3 ROBUSTNESS OF THE GABOR FILTER BANK FEATURES

In this part of the study the Gabor filter bank features are compared to several traditional feature extraction schemes in terms of robustness against extrinsic and intrinsic variability of speech. It is structured as follows: First, in Sec. 2.3.1 the traditional feature extraction schemes which serve as reference are introduced. Then, in Sec. 2.3.2 the experiments

used for evaluation are presented. Finally, the results are presented and discussed in Sec. 2.3.3.

2.3.1 Baseline features

Standard Mel-frequency Cepstral Coeffient (MFCC) (Davis and Mermelstein, 1980) features are used as a reference. MFCCs are calculated by applying a discrete cosine transform to spectral slices of the Mel-spectrogram. The coefficients, encoding the spectral envelope of quasi-stationary speech segments, are then used as features for ASR. The Rastamat toolbox for Matlab (Ellis, 2005) is used to generate 13-dimensional MFCC features (MFCCs), which resemble the features obtained with the HTK package (Young et al., 2001). Adding the first and second discrete derivative results in 39-dimensional features. As a second reference, cepstral mean subtraction (CMS), a blind deconvolution technique which Schwartz et al. (1993) found to improve recognition accuracy and robustness to changes of communication channel characteristics is applied to the MFCCs; these features are referred to as MFCC CMS. The baseline MFCC features on the Aurora 2 task from Pearce and Hirsch (2000) are referred to as MFCC WI007. As a third reference, 8th order Perceptual Linear Prediction (PLP) (Hermansky, 1990) features that have undergone additional modulation band pass filtering, are calculated with the Rastamat toolbox. The filtering emphasizes the relative differences between spectra, hence, these features are referred to as RASTA-PLP features (Hermansky and Morgan, 1994). RASTA-PLPs have been reported to be robust, especially in the presence of channel distortions (Hermansky and Morgan, 1994). The addition of delta and acceleration coefficients results in 27-dimensional feature vectors.

2.3.2 Experiments

In this section, the experimental setups that are employed to evaluate the robustness against extrinsic and intrinsic variability in speech are presented.

2.3.2.1 Effect of extrinsic factors (Aurora 2 and Numbers95)

For evaluation of robustness against extrinsic variability the Aurora 2 framework (Sec. 2.2.2.1) is used. Since several parameters of the GBFB

features were optimized with the Aurora 2 framework, additional experiments are performed with a different speech corpus and a different state-of-the-art back end. The aim of this experiment is to check whether the results for GBFB features on the Aurora 2 task translate to a different ASR setup without further adaptation. The speech database chosen was NUMBERS95 (Cole et al., 1995) that contains strings of spoken numbers collected over telephone connections. The data consists of zip codes and street numbers, extracted from thousands of telephone dialogues. In addition, this corpus contains data from male and female American-English speakers of different ages. Following the experimental setup from Zhao and Morgan (2008), the corpus was divided in a training set (with 3590 utterances which approximates to 3 h of data) and a testing set (1227 utterances or 1 h of data). There are two experimental conditions for the testing set; one contains all testing-set utterances in clean condition; the other contains the utterances in noise-added conditions. The noise-added test set is created using the principles delineated in the Aurora 2 task (Pearce and Hirsch, 2000) using noises of different signal-to-noise ratios from the NOISEX-92 collection (Varga and Steeneken, 1993).

Features were mean and variance normalized and used to train the GMM/HMM recognizer *Decipher* developed by Stanford Research International (SRI). This state-of-the-art system is used to compare spectro-temporal and other features against a competitive baseline. Gender-independent, within-word triphone HMM models were based on a phone model comprising 56 consonants and vowels. Parameters were shared across 150 states clustered with a phonetic decision tree, and a diagonal-covariance GMM with 16 mixture components modeled the observation distribution. Maximum Likelihood estimation was used to estimate the parameters. Features are used either as direct input to Decipher, or processed in a Tandem system (Hermansky et al., 2000) that uses a multi-layer perceptron (MLP) to estimate the phone posterior probabilities for each feature frame. The posteriors are then log-transformed and decorrelated with a principal component analysis, in order to match the orthogonality assumption of the HMM decoder. For experiments that employ MLP-processing, the training of the neural net was carried out with phonetically labeled digit sequences from Numbers95 training set. The phoneme labels were obtained from forced alignment. The MLP used 9 frames of temporal context which resulted in $9 \times 331 = 2927$ input units, 160 and 56 units were used for the hidden and output layer,

respectively. For the last set of experiments, 13-dimensional MFCC features with delta and double-delta features were appended to the MLP-transformed Gabor features, resulting in 71-dimensional feature vectors, since this has been reported to increase accuracies in other research that used spectro-temporal features as input to ASR (Zhao and Morgan, 2008). The results for the MFCC, MFCC CMS, RASTA-PLP and GBFB features are presented, compared and discussed in Sec. 2.3.3.1 on the Aurora 2 task, and in Sec. 2.3.3.2 on the Numbers95 task.

2.3.2.2 Effect of intrinsic factors (OLLO framework)

To evaluate the robustness against intrinsic variability in speech, an experimental framework that aims at the analysis of factors such as speaking style, effort, and rate is proposed. In this framework the sensitivity of different feature types against such variabilities is evaluated by performing experiments with a mismatch between the training and test data. The degradation in performance quantifies the robustness against a specific mismatch. A statistical test, McNemar's Test as suggested by Gillick and Cox (1989), is employed to test the results for significant differences between the feature types.

The speech database used for this framework is the Oldenburg Logatome Corpus (OLLO) (Wesker et al., 2005), which consists of nonsense vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) logatomes with identical outer phonemes (e.g., [p u p] or [a p a]). The database contains 150 different logatomes (70 VCVs and 80 CVCs), spoken by German speakers in different speaking styles. During the recordings, the speakers were asked to produce the utterances normally, with varied speaking effort (loud and soft speaking style), varied speaking rate (fast and slow), and with rising pitch, which is referred to as category "questioning". Three repetitions of each logatome in each speaking style were collected in order to obtain a sufficient amount of ASR training data. This resulted in 150 (logatomes) $\times$ 6 (speaking styles) $\times$ 3 (repetitions) = 2700 utterances per speaker.

For the OLLO framework that we propose to evaluate robustness against intrinsic variability in speech, speech data from ten speakers without dialect is used. Six training and six test conditions are defined, which correspond to the various speaking styles contained in the OLLO corpus (fast, slow, loud, soft, questioning and normal). Training and testing on each condition resulted in 36 individual experiments. The

experiments are carried out using a 10-fold cross validation, i.e., speech signals of nine speakers are used for training, and the data of the remaining speaker is used for testing. This procedure is repeated for all speakers, and the individual scores are averaged.

As for the Aurora 2 framework, results for MFCC features serve as baseline. These are fed to an HMM using HTK (Young et al., 2001). The HMM is configured as word recognizer, i.e., the classification task is to make a 1-out-of-150 decision based on a dictionary that contains the transcription of the 150 logatomes. The number of HMM states per logatome is set to 16, which was found to be the optimal value in pilot experiments for MFCC features. Other parameters, such as the increase of Gaussian mixtures during training, are copied from the Aurora 2 setup. Additionally, performance of MFCC features with CMS and RASTA-PLP features is evaluated. Since the PLP part of this algorithm accounts for the reduction of speaker-dependent information it is interesting to see whether it improves the robustness against intrinsic variability. The results are presented in Sec. 2.3.3.3.

2.3.3 Results and discussion

2.3.3.1 Robustness against extrinsic variability (Aurora 2)

This section presents the results of recognition experiments with GBFB, MFCC, MFCC CMS and RASTA-PLP features that are carried out with the aim of quantifying the robustness against extrinsic variability (additive noise and channel distortions) on the Aurora 2 task (employing the HTK recognizer).

Absolute results for the various feature types are presented in Table 2.5. In terms of average word recognition accuracies (WRAs) GBFB features outperform MFCC and RASTA-PLP features with clean (multi) condition training by 8 (1) percentage points and 2 (3) percentage points, respectively. With cepstral mean subtraction MFCC features achieve a slightly higher average WRA than GBFB features. The overall relative improvement over MFCC WI007 standard features, which is calculated as described in Sec. 2.2.2.1 is presented in Table 2.6. GBFB features improve the WER of standard MFCC WI007 features by more than 16% on average with multi condition training and by 28% on average with clean condition training. The use of MFCCs with CMS improves the baseline by 12% on average with multi condition training and by 27%

Table 2.5: *Recognition accuracies in percent for GBFB, MFCC WI007, MFCC CMS, and RASTA-PLP features on the Aurora 2 task for different noise conditions, average word recognition accuracies for each test set and standard deviation over all noise conditions. The average values presented here are obtained by averaging over SNRs from* 0 *dB to* 20 *dB.*

		Test set A				Test set B				Test set C			
		Sub.	Bab.	Car	Exh.	Res.	Str.	Air.	Sta.	Sub. m	Str. m	Avg.	rms
GBFB	mul.	89.0	88.0	86.1	88.1	90.0	88.2	90.7	85.9	88.8	86.3	88.1	**1.62**
	avg.		87.8				88.7			87.6			
	cle.	70.9	67.0	60.0	64.3	69.2	64.5	68.9	65.0	68.8	63.4	66.2	3.33
	avg.		**65.6**				66.9			**66.1**			
RPLP	mul.	87.5	84.6	83.7	83.7	83.8	84.5	85.6	82.1	87.1	84.3	84.7	1.64
	avg.		84.9				84.0			85.7			
	cle.	64.3	63.1	59.5	59.9	66.3	63.8	66.4	63.1	64.5	63.7	63.5	**2.30**
	avg.		61.7				64.9			64.1			
MFCC	mul.	89.1	88.4	86.8	80.0	86.6	87.8	88.3	86.2	83.5	85.7	87.0	1.67
	avg.		88.1				87.2			84.6			
	cle.	66.7	47.8	58.1	62.3	50.0	60.7	49.6	53.1	65.3	66.7	58.1	7.40
	avg.		58.7				53.4			66.0			
MFCC^{CMS}	mul.	90.3	89.8	84.9	88.0	90.0	87.9	91.1	86.7	89.9	87.9	**88.7**	1.92
	avg.		**88.2**				**88.9**			**88.9**			
	cle.	64.1	67.7	62.2	62.8	71.3	65.6	72.0	67.7	64.2	65.4	**66.3**	3.35
	avg.		64.2				**69.2**			64.7			

on average with clean condition training. When concatenating GBFB features with MFCC features from the Rastamat toolbox with CMS applied, a further improvement of a few percent is achieved, indicating that these feature types carry complementary information. RASTA-PLP features outperform the standard MFCC WI007 features by about 14% with clean condition training, but with multi condition training they perform 31% worse than MFCCs.

The relative improvements over the baseline for the test sets A, B, and C are also presented in Table 2.6. In addition to the reference features, the performance of GBFB features concatenated with MFCC CMS features is shown. GBFB features outperform the MFCC WI007 baseline in all test conditions (test sets A, B and C). For multi-condition training, the relative improvements for test set A and test set B are comparable

Table 2.6: *Relative reduction of the word error rate obtained with GBFB, MFCC CMS, RASTA-PLP features, and with GBFB features concatenated with MFCC CMS features compared to the MFCC WI007 baseline for the test sets A, B and C.*

		Test set A 4 conditions	Test set B 4 conditions	Test set C 2 conditions	Average over all conditions
GBFB	clean	22.9	40.6	15.1	28.4
	multi	11.4	15.2	27.7	16.1
MFCC^{CMS}	clean	15.9	**45.5**	10.9	26.7
	multi	3.6	16.1	19.0	11.6
RPLP	clean	2.6	31.4	4.0	14.4
	multi	−33.3	−40.0	−12.7	−31.9
GBFB $\&\text{MFCC}^{\text{CMS}}$	clean	**26.3**	43.8	**18.0**	**31.6**
	multi	**16.7**	**24.4**	**33.0**	**23.0**

with improvements of about 13%, which indicates that GBFB features generalize as well as MFCC features with respect to mismatches in noise types when training with noisy data. For test set C, the relative improvement for GBFB features is more distinctive with about 28%. MFCC features with CMS improve the WI007 baseline in test set B and C, i.e., when noise or communication channel characteristics changed compared to the training data. The improvements with test set C (channel distortions) for MFCCs with CMS is smaller than with GBFB features. RASTA-PLP features perform worse than the MFCC WI007 baseline with multi condition training on all test sets. For test set C (channel distortions), the difference between RASTA-PLP and MFCC features is smaller than on test set A and B.

For clean training, the differences between GBFB and MFCC WI007 features are larger compared to multi-condition training with a relative decrease of the WER for GBFB features in the range of 15% to 40%. With MFCC CMS features and clean condition training the improvements are also larger compared to the multi condition training. The smaller relative improvements in test set C are a result of the relatively high performance of the MFCCs (cf. Table 2.5). RASTA-PLP are consistently better than the baseline for clean condition training, but do not improve results with multi-condition training.

The standard deviation of recognition scores for various noise conditions is reported as a measure for the stability of scores in the last column of Table 2.5. The results for clean condition training are of special interest in this case, since they can be interpreted as the robustness in the presence of unknown noise sources. The standard deviation for MFCCs (7.4 percentage points) is approximately twice as high as for GBFB and three times as high as for RASTA-PLP features (3.4 percentage points and 2.3 percentage points, respectively). This indicates that GBFB and RASTA-PLP features are less sensitive to mismatches between training and test data than MFCC features. When applying CMS to the MFCC features the standard deviation decreases to the level of GBFB features. For multi-condition training, the standard deviations are smaller than 2 percentage points, with only small differences between the feature types but MFCCs with CMS, which show a slightly higher standard deviation.

A comparison of the relative improvements of GBFB features over MFCC features in Table 2.6 with the absolute results in Table 2.5 shows that the differences between both in terms of WRAs is rather small. This suggests that the improvements of GBFB over MFCC features are obtained at high SNRs. This is investigated further by separating the WRA results by SNRs. The average WRAs for each feature type and for each SNR are depicted in Fig. 2.8. The ordinate is scaled as a logarithmic error axis and labeled with the corresponding WRA. The distance of two horizontal lines corresponds to a halving/doubling the WER so that the results in terms of relative improvements are projected linearly, i.e., they are proportional to the relative improvement of the averages over all noise conditions.

For all feature types, a strong decrease of the WRA is observed when the noise level is raised, with 95% WRA for clean utterances to down to scores below 30% at an SNR of $-5\,\mathrm{dB}$. The major differences between the feature types are observed at high SNRs (20 dB to 5 dB). For clean training, the decrease in performance is more pronounced than for multi-condition training. While using noisy training data and testing with clean utterances results in lower scores compared to clean training, the overall performance (tested over multiple SNRs and noise types) is improved with multi-condition training as expected. When training with clean utterances, RASTA-PLP features outperform the MFCC WI007 baseline at almost all SNRs, which confirms the observation that RASTA-PLPs are more robust than short-term spectrum based features in unknown noise

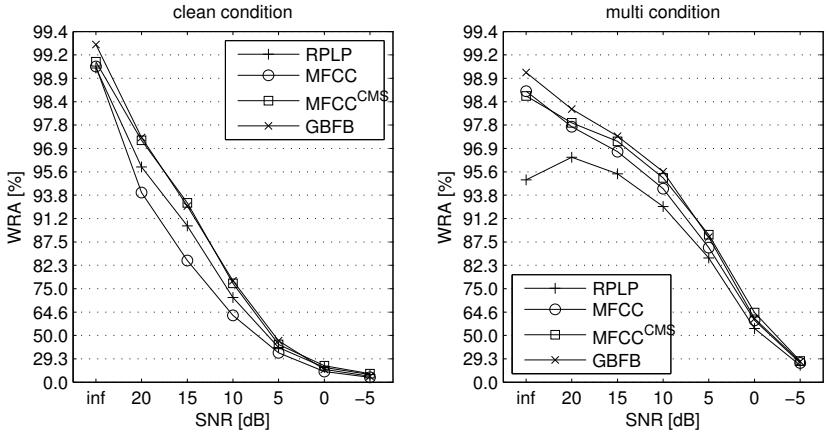

Figure 2.8: *Recognition accuracies in percent for GBFB, MFCC and RASTA-PLP features at different test SNRs for multi and clean condition training on the Aurora 2 task. The ordinate is a logarithmically scaled WER-axis and labeled with the corresponding WRA. The distance of two horizontal lines corresponds halving/doubling the WER.*

conditions (Hermansky and Morgan, 1994). However, for multi-condition training, which allows the ASR system to adapt to different noise types, the MFCCs produce higher scores than RASTA-PLPs. GBFB features improve the scores of MFCC WI007 and RASTA features at almost all SNRs: The robustness against additive noise is found to be higher than for RASTA features over a wide range of test conditions (i.e., clean signals and SNRs from 5 dB to 20 dB), and additionally are found to outperform the MFCC WI007 baseline for multi-condition training in these test conditions. MFCCs with CMS and GBFB features perform similarly well. However, when testing on clean or low-noise signals with multi-condition training, GBFBs outperform all feature types.

The average relative improvement of GBFB, MFCC CMS and RASTA-PLP over MFCC WI007 features depending on the SNR (averaged over all noise conditions) is depicted in Table 2.7. The results are comparable to those presented in Fig. 2.8. GBFB features outperform MFCCs at all SNRs. The best improvements are obtained at SNRs above 0 dB SNR, while at low SNRs the differences are negligible. While RASTA-PLP

Table 2.7: *Relative reduction of the word error rate obtained with GBFB and RASTA-PLP features compared to the MFCC baseline (averaged over all noise conditions). Values in the column* average *are averaged over SNRs from* 20 *dB to* 0 *dB.*

SNR [dB]		∞	20	15	10	5	0	−5	Avg.
GBFB	mlt	23.9	23.4	21.0	21.3	13.5	1.7	1.8	16.2
	cln	27.0	45.1	45.1	34.9	14.3	2.8	1.5	28.4
RASTA-	mlt	−274.6	−56.4	−38.5	−31.8	−18.4	−14.5	−3.7	−31.9
PLP	cln	−3.8	16.9	28.4	16.7	4.7	5.3	3.9	14.4
MFCC^{CMS}	mlt	−7.6	5.3	13.4	13.0	16.3	10.1	2.9	11.6
	cln	5.5	42.2	45.9	29.9	8.0	7.6	5.2	26.7

features outperform MFCCs with clean training at all SNRs they do not improve MFCC results when testing on clean data. The relative improvements for low SNRs (0 dB, −5 dB) are higher than with GBFB features but still below 6%. This means that RASTA-PLP and GBFB features are more robust than MFCCs when the noise signal energy still is about 5 dB below the level of the speech signal energy. When learning the noise characteristics (multi-condition), GBFB features perform better and RASTA-PLP features perform worse than MFCCs, with the greatest differences in relative improvements at high SNRs. MFCCs with CMS improve the MFCC WI007 baseline at almost all SNRs, with the single exception of multi-condition training and clean testing. When testing on clean data GBFB features improve the MFCC baseline by more than 6%, while the baseline was not improved with the other feature types. For testing on clean speech data, GBFB features improve the baseline by about 25%.

A 28% relative improvement of GBFB features over the MFCC baseline is observed when the channel characteristics of training and testing differ. We assume that MFCCs are stronger affected by such influences since the spectrogram is integrated over the full bandwidth, which might be a disadvantage compared to the localized GBFBs. Cepstral mean subtraction seems to alleviate this disadvantage, but not to the extent that was observed for GBFB features. Further, considering even higher frequencies (above 4 kHz) could be beneficial with the Gabor filter bank features. While the MFCC features would change fundamentally, for the

Gabor filter bank features it would mean an extension to more center frequencies. This should be evaluated on a suitable task in the future.

GBFB features were shown to perform better in the high SNR range from 20 dB to 5 dB than MFCC and RASTA-PLP features and equally well at lower SNRs (0 dB and −5 dB) on the Aurora 2 task, which evaluates robustness of ASR systems against extrinsic variability. GBFB features also slightly outperform MFCC features with CMS. This suggests that the physiologically inspired representation of speech signals by GBFB features is more robust to extrinsic variability than those of MFCCs and RASTA-PLPs over a wide range of SNRs and is similarly robust to extrinsic variability as MFCC with CMS. Further, improvements of about 25% for testing on clean data are observed which points out the beneficial effect of spectro-temporal information on feature level.

2.3.3.2 Robustness against extrinsic variability (Numbers95)

In this section the results for the NUMBERS recognition task, which was conducted with the aim of checking whether the results from the Aurora 2 task translate to a different ASR setup, are presented. Absolute and relative results obtained on the NUMBERS recognition task with the SRI Decipher recognizer are shown in Table 2.8. In this scenario, MFCC and MFCC CMS features perform best, while for GBFB and RASTA-PLP features relatively high error rates are observed. A possible reason may be that GBFB features encode up to 400 ms context and RASTA-PLP features up to 200 ms context, and may thus not be suited as well as the MFCCs (up to 100 ms context) for triphone based models.

We then tested if mapping the features to phoneme posteriors, which we assume to be suitable to build phone base models, by means of a multi-layer perceptron (MLP) improves the recognition performance. This MLP processing, which was reported to improve results in earlier studies (Hermansky et al., 2000), almost halved the error rate of GBFB features without MLP processing for clean testing and also improved the results with RASTA-PLP features, but the performance was still below the baseline. Using MFCCs and MFCCs with CMS in conjunction with MLP processing leads to small improvements when testing on noisy data, but not for testing on clean data. The results with "long-term context" features, i.e., GBFB features and also RASTA-PLP to a smaller extent, improved much more by the MLP processing than the results with the already well performing "short-term context" features. Another

Table 2.8: *Word error rates for the NUMBERS95 task with SRI's ASR system Decipher. Features were either used as direct input to the classifier, processed with an MLP, or first MLP-processed and then concatenated with a different feature vector.*

		Absolute WER		Rel. imp.	
	Feat. dim.	Clean	Avg. noisy	Clean	Avg. noisy
MFCC	39	3.7	19.4	–	–
MFCC_{CMS}	39	3.7	17.8	1.1	8.1
RASTA-PLP	27	6.0	23.2	−59.6	−19.7
GBFB	311	9.1	22.9	−142.3	−16.2
MLP (MFCC)	32	4.0	19.2	−7.9	1.0
MLP (MFCC_{CMS})	32	3.7	16.9	1.1	12.7
MLP (RASTA-PLP)	32	5.7	20.8	−51.1	−7.6
MLP (GBFB)	32	4.6	19.9	−21.9	−2.6
MLP (GBFB) and MFCC	71	3.3	16.8	10.7	13.5
MLP (GBFB) and MFCC_{CMS}	71	3.2	16.6	15.2	14.1

reason for the high error rate with GBFB features may be the high dimensionality of the features. While the GMM/HTK back end of the Aurora 2 framework had no problems with high dimensional features, the Decipher recognizer may be tuned to the dimensionality of typical feature types, hence performing better with the low dimensional MLP processed GBFB features.

The improvements over the MFCC baseline that were observed on the Aurora 2 task with GBFB features do not translate directly to setups with different back ends. This is because the back end imposes strong restrictions upon the statistical characteristics of the used features. These restrictions depend on many factors like the training material, the acoustic model type, and the complexity of the recognizer. We assume that the shorter triphone models of the Decipher back end favor features with less temporal context compared to the whole word models of the HTK recognizer on the Aurora 2 task. However, adapting the features to the restrictions of the back end improves the recognition performance. GBFB features are long-term context features and seem to work better with models that can make full use of long-term context (word models).

The fact that error rates were lower when combining MFCC and GBFB features for the Aurora 2 task motivated a combination of MLP-processed features with MFCCs. For this setup, the MFCC baseline is outperformed by more than 10% (for combinations with MFCCs), and $14 - 15\%$ for combinations with MFCC CMS. We also tested other combinations (such as MLP-processed spectral features that are combined MFCCs); however, none of these yielded results above the baseline.

With the Decipher back end, the baseline was not improved when only using GBFB features, but when using the features in a Tandem system and combining them with spectral features, the baseline was outperformed by $14 - 15\%$. This result confirms earlier studies that reported an increase of the robustness of ASR system against additive noise and channel distortions when using MLP-processed spectro-temporal features in conjunction with concatenated MFCCs (Meyer and Kollmeier, 2011a; Zhao et al., 2009). It also supports the hypothesis that MFCCs and GBFB features encode complementary information that is useful for robust ASR.

2.3.3.3 Robustness against intrinsic variability

This section presents the results of recognition experiments with GBFB and baseline features that are carried out with the aim of quantifying the robustness against variability due to intrinsic sources (arising from variation in speaking rate, effort and style). The ASR task is to classify VCV and CVC utterances from the OLLO database, as described in Sec. 2.3.2.2. The absolute word recognition accuracies are depicted in Table 2.9. Scores are presented for each combination of training and test speaking styles, which results in 6×6 individual scores per feature type. RASTA-PLP and MFCC CMS features produce almost consistently worse scores than MFCC and GBFB features and are therefore not included in Table 2.9.

When averaging over all scores obtained for mismatched training and test conditions (off-diagonal elements in Table 2.9), the recognition scores for GBFB and MFCC features are very similar with 59.3% and 58.8%. RASTA-PLPs produce an average of 55.6% (not shown) and MFCC CMS features produce an average of 56.4% (also not shown). All feature types exhibit similar error patterns, which are depicted in Fig. 2.9. Not surprisingly, the best scores are obtained with matched condition training. Compared to matched train-test conditions, the word

Table 2.9: *Absolute WRA in percent for GBFB and MFCC features on the OLLO logatome recognition task.* ***Averages are calculated over mismatched conditions****. Matched conditions are printed in italics and are not considered for averages.*

	Train\Test	Fast	Slow	Loud	Soft	Quest.	Normal	Average
GBFB	Fast	*74.0*	47.0	62.7	52.6	49.5	72.7	56.9
	Slow	45.6	*76.7*	46.5	66.3	39.4	69.9	53.5
	Loud	70.3	56.3	*78.7*	47.7	50.5	75.6	60.1
	Soft	51.9	64.0	42.9	*74.1*	49.1	67.7	55.1
	Questioning	61.8	65.7	56.0	68.0	*78.3*	74.9	65.3
	Normal	70.7	65.7	64.4	66.5	56.0	*81.4*	64.7
	Average	60.1	59.7	54.5	60.2	48.9	72.2	59.3
MFCC	Fast	*72.4*	52.2	60.8	49.7	49.4	72.3	56.9
	Slow	51.3	*77.5*	50.5	66.2	50.0	73.1	58.2
	Loud	68.6	58.9	*77.1*	43.1	52.5	72.7	59.2
	Soft	50.1	62.7	31.9	*71.0*	45.0	64.1	50.8
	Questioning	61.0	66.0	54.8	63.9	*76.8*	72.2	63.6
	Normal	70.6	68.4	61.3	64.4	56.4	*79.9*	64.2
	Average	60.3	61.6	51.8	57.5	50.6	70.9	58.8

recognition accuracies of the mismatched conditions show a degradation of about 17 percentage points on average. For MFCCs, the category "normal" for training yields the highest scores when considering the average over all six test conditions. For GBFB features, the category "questioning" for training yields slightly better (0.6 percentage points) word recognition accuracies than the category "normal". When using normally spoken utterances for the training, the reduction in WER is roughly -70% when testing on mismatch conditions (average over all feature types).

For the chosen order of sources of variability in Fig. 2.9, a checker board pattern is observed in the upper left part of each matrix. Relatively high accuracies are obtained for the training-test pairs (fast, loud) and (slow, soft), which indicates that utterances from these categories share properties that are embedded in the acoustic models of the HMM during training. On the other hand, the pairs (fast, slow), (fast, soft), (loud, soft) and (loud, slow) yield a score that is degraded by about 24

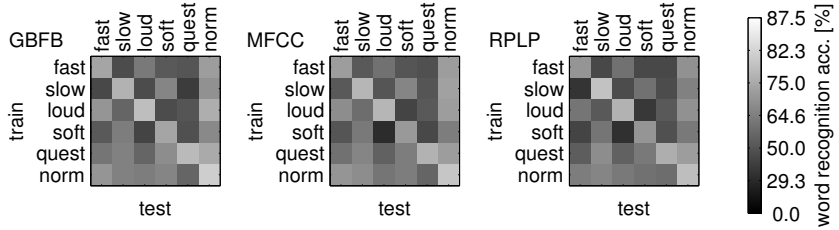

Figure 2.9: *Logarithmic word error rates for different training and testing conditions on the OLLO logatome recognition task. The colorbar indicates the corresponding word recognition accuracies. (left) GBFB features; (middle) MFCC features; (right) RASTA-PLP features.*

Table 2.10: *Relative improvement of GBFB features over MFCC features in percent. Scores for matched conditions (diagonal elements of the table, printed in italics) are not considered for the average values.*

Train\Test	Fast	Slow	Loud	Soft	Quest.	Normal	Average
Fast	*+6*	−11	+5	+6	+0	+2	+0
Slow	−12	*−4*	−8	+0	−21	−12	−11
Loud	+5	−6	*+7*	+8	−4	+11	+3
Soft	+4	+4	+16	*+11*	+8	+10	+8
Quest.	+2	−1	+3	+11	*+6*	+10	+5
Normal	+0	−9	+8	+6	−1	*+8*	+1
Average	−0	−5	+5	+6	−4	+4	+1

percentage points on average (average over all feature types) compared to the respective matched condition.

While GBFB and MFCC features perform similarly well on average, a detailed analysis of the recognition results with respect to speaking rate, style, and effort reveals systematic differences. Figure 2.10 shows in which particular conditions the differences between MFCC features and GBFB features are significant, i.e., the p-values are less than 0.01 according to McNemar's Test as proposed by Gillick and Cox (1989). The differences in terms of relative improvement of WER are depicted in Table 2.10. Only the mismatch conditions (off-diagonal elements) are considered for the average. These values can be interpreted as the sensitivity of GBFB

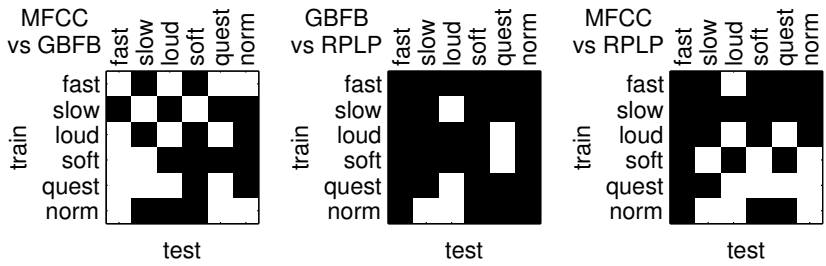

Figure 2.10: *Analysis of differences between the feature types according to McNemar's Test. Black: Significant differences with $p < 0.01$; white: Not significant. (left) Differences between GBFB and MFCC features; (middle) differences between GBFB and RASTA-PLP features; (right) Differences between MFCC and RASTA-PLP features.*

features (compared to MFCC results), or the robustness against intrinsic variability.

The results show that on average GBFB features are slightly more sensitive against such mismatches (with a 0.2% relative degradation when averaging over all combinations of training and testing). The relative reduction of the WER with GBFB features compared to MFCCs shows that MFCCs exhibit a better recognition performance for high and low speaking rate (categories "fast" and "slow"), while GBFB features are better suited when the talker changes his speaking effort (categories "loud" and "soft"). This trend is consistent both for training and for testing. Interestingly, when the recognizer is trained with utterances with rising pitch ("questioning"), GBFB feature perform better than MFCC features (row "questioning" in Table 2.9). On the other hand, when testing is performed with logatomes spoken as question, this results in higher scores with MFCC features than with GBFB features (column "questioning" in Table 2.10).

On average, the performance with MFCC features deteriorates by about 70% (relative improvement calculated as explained in Sec. 2.2.2.1), the GBFB features' performance drops by about 80% when training and test data categories mismatch. Compared to MFCCs, GBFB features seem to perform similarly well on average in the tested mismatching conditions of intrinsic sources of variability. However, they appear to be

slightly more susceptible to such variations than MFCCs, since they tend to perform better in matched conditions, which are not considered for averages.

In the presence of intrinsic variation (measured with the OLLO recognition task, cf. Sec. 2.3.2.2) considerable degradations are observed for all feature types. Compared to the matched condition scores, the average relative increase of the word error rate is between 70 and 85% (for MFCC and MFCC CMS features, respectively). In order to analyze the robustness against intrinsic factors, the scores obtained with mismatched training are of special interest. In the presence of variation caused by intrinsic sources, GBFB and MFCC features exhibit a comparable overall performance. However, when individual sources of variability are considered, the error patterns for both feature types show statistically significant differences, indicating that these feature types carry — at least to some extent — complementary information.

Using mismatched conditions in training and testing shows that the training-test pairs (fast, loud) and (slow, soft) produce relatively high accuracies. This trend is observed for all feature types. The combinations (fast, soft) and (slow, loud) on the other hand produce rather low scores. It may be that these categories share several acoustic properties, since speakers, e.g., unconsciously increase their speaking effort when asked to produce an utterance with high speaking rate. Such an interaction might also explain the high scores for the pair (soft, slow).

GBFB features are observed to perform better than MFCCs when training on utterances pronounced with rising pitch (category "questioning"), but worse when testing on utterances of this category. A possible explanation for this observation is that GBFB features can account for spectral details such as pitch information. However, to account for the larger variability caused by changes in pitch, the according speech data has to be included in the training material.

For matched training and test conditions, the best average results are obtained with GBFB features. However, GBFB features are not found to be more robust than MFCC features against intrinsic variability, i.e., spectro-temporal information does not seem to improve robustness against intrinsic variability in general. The differences observed between the feature types indicate that the information captured in the feature calculation process is at least partially complementary; hence, the combi-

nation of these features (e.g., in a multi-stream framework) could result in an improvement of the ASR performance.

For RASTA-PLP and MFCC CMS features, relatively low scores are obtained. The fact that both the training and testing with the OLLO database are performed with clean utterances might explain this observation for RASTA-PLP, since the Aurora 2 experiment showed that these features only improved the baseline for additive noise and channel distortions. Moreover, the calculation of RASTA-PLPs includes temporal filtering, which might be suboptimal for very short utterances such as the phoneme combinations used for the OLLO corpus, although GBFB features also capture temporal information to a comparable extent. For CMS, an integration over the whole utterance is needed. Maybe the shortness of the utterances does not allow for a good estimation of the mean value, thus resulting in a mismatch that deteriorates performance.

2.4 SUMMARY AND FURTHER DISCUSSION

2.4.1 Robustness of GBFB features against extrinsic variability

The performance of a robust speech recognition system depends on the interaction of its parts. The results presented in this study show that improvements over a MFCC baseline can be obtained with physiologically inspired spectro-temporal features when the back end's assumptions about the statistical feature characteristics are met. It can be assumed that the properties of the Gabor filter bank result in a filter output with limited redundancy between individual components and mostly independent features with up to 400 ms of temporal context. Depending on the task and the back end it may be favorable to apply MLP processing to the GBFB features in order to meet the back end's assumptions about the features. In this case improvements over the unprocessed GBFB features can be expected, but not necessarily an improvement of a MFCC baseline.

2.4.2 Complementary information

The experiments show that the combination of MFCC CMS and GBFB features, possibly processed with an MLP, results in a further increase of recognition performance. Presumably, there most possibly is a part

of information important for ASR represented in a better suited form by MFCCs than by GBFB features and vice versa. This also means that neither MFCC nor GBFB features are sufficient to extract all the characteristics of human speech.

Earlier studies using spectro-temporal features for ASR presented evidence that MFCCs and spectro-temporal features carry complementary information (Meyer and Kollmeier, 2011a; Zhao et al., 2009). This finding is also supported by the experiment that analyzed the sensitivity against intrinsic variation, since the performance obtained with MFCC and GBFB features significantly differ in many conditions. For example, cepstral features are found to be better suited for recognition of fast and slowly spoken utterances, while GBFB features produce better results when the speaking effort is varied. The results of the multi-stream experiment carried out on the Aurora 2 task, which improves performance over GBFB and MFCC CMS features by concatenating them also supports this finding.

2.4.3 Future work

The Gabor filter bank could be used for speech analysis in order to evaluate the importance of modulation frequencies: The integration of the outputs of the localized Gabor filters results in a spectro-temporal representation resembling the original spectrogram. When isolated spectro-temporal components are removed from the filter bank, their contribution to speech recognition may be assessed in tests with human listeners (resembling the ASR band-stop experiments carried out in this study).

The results investigating intrinsic variation of speech show that spectro-temporal and purely spectral ASR features produce significantly different results depending on the specific source of variability. Further, small improvements over using GBFB features are achieved when combining them with MFCC CMS features. Based on these observations, it may be worthwhile to further investigate methods to combine information from different feature streams, thereby exploiting the complementary information of the feature types. The output of the Gabor filter bank also contains purely spectral output, which may not be required (or even detrimental) when combined with MFCC features, which may also be subject of future investigations. Alternatively, the purely spectral output of the GBFB might be modified to closely resemble the extraction of

cepstral features, which would effectively integrate the informational content of MFCCs into Gabor features.

The GBFB features extend naturally to higher frequency bands. It should be evaluated if this behavior has an advantage over MFCC features that always project the whole bandwidth of the log Mel-spectrogram.

The parameters of the Gabor filter bank (i.e., the optimal number of oscillations under the envelope) are optimized on the Aurora 2 digit recognition task, but also show good performance on the OLLO logatome recognition task. However, when changing the back end, the GBFB features do not necessarily meet the assumptions made about them and can perform worse than traditional features. In this case the robustness of these systems may be improved by processing the GBFB features with a MLP and concatenating MFCC features. This suggests that the proposed GBFB features, possibly with MLP processing, may be applicable to a wider range of ASR recognition tasks with the same parameters, which should be assessed in future experiments. To further validate the findings, GBFB features should be tested on a large vocabulary speech recognition task.

It seems that not all of the 311 filter outputs extract useful information. Especially the highest spectro-temporal modulation frequencies seem to have a negative effect on the recognition performance. It could be that covering a rectangular region of the modulation domain is not optimal. Hence, feature selection techniques could further improve the performance.

2.5 CONCLUSIONS

The most important findings of this work can be summarized as follows:

— The use of spectro-temporal Gabor filter bank (GBFB) features increases the robustness of ASR systems against additive noise and mismatches of channel transmission characteristics (i.e., *extrinsic* sources of variability) compared to MFCC and RASTA-PLP features. For this, it can be necessary to process the GBFB features with a multi-layer perceptron (MLP) and combine them with MFCCs depending on the task and the back end. A MFCC baseline was also improved for high SNRs and clean speech. With a standard GMM/HMM recognizer, improvements of over 40% with clean training and over 20% with multi training were observed when the GBFB features were used as

direct input to the classifier. A state-of-the-art baseline system was outperformed by $14 - 15\%$ when GBFB features were first processed with a MLP and then combined with MFCC features. These findings indicate that the proposed feature extraction scheme results in a good representation of speech signals for ASR tasks. GBFB and MFCC features were found to extract partly complementary information regarding extrinsic and intrinsic sources of variability, which may be exploited in feature stream experiments.

— On average, MFCC and GBFB features are similarly affected by *intrinsic* variability of speech, while for RASTA-PLP features and MFCCs with CMS higher degradations are observed. When analyzing train-test pairs with unmatched intrinsic variations, the MFCC and GBFB scores show significant differences, which shows that the feature types exhibit different strength and weaknesses with respect to intrinsic factors.

— The analysis of specific modulation frequencies for ASR with GBFB features shows that temporal modulation frequencies from 6 Hz to 25 Hz and spectral modulation frequencies from 0.03 cycles/channel to 0.25 cycles/channel are important for robust speech recognition. Besides the information about the input level, spectral modulation frequencies of about 0.25 cycles/channel were found to be especially important for robust speech recognition. When using spectro-temporal features for ASR, the usable temporal modulation frequencies are shifted to higher frequencies than reported in the literature that analyzed spectral and temporal information separately.

ACKNOWLEDGMENTS

Supported by the DFG (SFB/TRR 31 “The active auditory system”). Bernd T. Meyer’s work is supported by a post-doctoral fellow-ship of the German Academic Exchange Service (DAAD). We would like to thank Jörg-Hendrik Bach and Jörn Anemüller for their support and contribution to this work. Thanks also to Suman Ravuri and Andreas Stolcke for providing support for the experiments with SRI’s recognition system, and the two anonymous reviewers, who made valuable suggestions to improve this study.

3 Normalization of spectro-temporal Gabor filter bank features for improved robust automatic speech recognition systems

ABSTRACT

Physiologically motivated feature extraction methods based on 2D-Gabor filters have already been used successfully in robust automatic speech recognition (ASR) systems. Recently it was shown that a Mel Frequency Cepstral Coefficients (MFCC) baseline can be improved with physiologically motivated features extracted by a 2D-Gabor filter bank (GBFB). Besides physiologically inspired approaches to improve ASR systems technical ones, such as mean and variance normalization (MVN) or histogram equalization (HEQ), exist which aim to reduce undesired information from the speech representation by normalization. In this study we combine the physiologically inspired GBFB features with MVN and HEQ in comparison to MFCC features. Additionaly, MVN is applied at different stages of MFCC feature extraction in order to evaluate its effect to spectral, temporal or spectro-temporal patterns. We find that MVN/HEQ dramatically improve the robustness of MFCC and GBFB features on the Aurora 2 ASR task. While normalized MFCCs perform best with clean condition training, normalized GBFBs improve the ETSI MFCCs features with multi-condition training by 48%, outperforming the ETSI advanced front-end (AFE). The MVN, which may be interpreted as a normalization of modulation depth works best when applied to spectro-temporal patterns. HEQ was not found to perform better than MVN.

This chapter is a reformatted reprint of "Normalization of spectro-temporal Gabor filter bank features for improved robust automatic speech recognition systems", M. R. Schädler and B. Kollmeier, which was published in the proceedings of INTERSPEECH 2012 pp. 1812–1815. Reprinted by permission of the publisher. The original article can be found at http://www.isca-speech.org/archive/interspeech_2012/i12_1812.html.

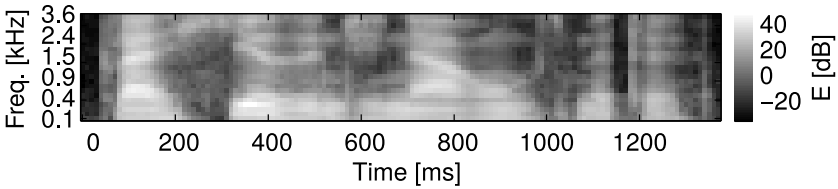

Figure 3.1: *Logarithmically scaled Mel-spectrogram of speech. Light areas denote high energy. The representation of speech through a log Mel-spectrogram is an element of many feature extraction algorithms for robust ASR systems.*

3.1 INTRODUCTION

After decades of research in the area of automatic speech recognition (ASR) still no system exists that would equal humans ability to recognize speech. Especially in acoustically adverse conditions (background noise, spectral coloring, reverberation) there is a big gap in performance of about 15 dB between humans and machines. Tackling the long-term goal to improve the robustness of ASR systems to the level of humans, several approaches exist. One approach is to mimic the signal processing of the human auditory system or rather, to integrate its principles in terms of effective models into ASR systems. This proved to work for the well known part of the auditory system as today many robust ASR systems employ features based on a logarithmically scaled Mel-spectrogram like the one depicted in Fig. 3.1. This representation of speech roughly reflects the frequency selectivity and the compressive loudness perception of the human ear. Beyond the log Mel-Spectrogram there were several successful attempts to integrate single auditory principles, like the extraction of physiologically motivated (Qiu et al., 2003) spectro-temporal patterns, into an ASR system to improve its robustness (Kleinschmidt and Gelbart, 2002a). The early spectro-temporal features used additional processing with neural nets to improve a MFCC baseline. Recently, a filter bank of spectro-temporal filters which extracts features that can be used directly with GMM/HMM recognizers and improved a MFCC baseline was presented (Schädler et al., 2012a). But generally, the use of the most detailed models of the auditory system does not result in the most robust ASR systems. One reason for this might be that the use of GMM/HMM

based back-ends entrains certain restrictions on the feature characteristics. A different approach is therefore the use of statistical methods to better match the requirements of state-of-the-art GMM/HMM based back-ends. Normalization techniques like MVN (Viikki and Laurila, 1998) or HEQ (De La Torre et al., 2005) have shown to improve the robustness of systems based on traditional MFCC features. In this study both approaches are combined and normalization methods are applied to the physiologically motivated spectro-temporal Gabor filter bank (GBFB) features in comparison to traditional MFCC features. Further, the effect of MVN/HEQ is interpreted as a normalization of modulation depth and its effect on temporal, spectral, and spectro-temporal patterns is investigated.

3.2 METHODS

3.2.1 Gabor filter bank features

The Gabor filter bank (GBFB) features are based on a log Mel-spectrogram with 23 Mel-bands between 64 Hz and 4 kHz, 10 ms window shift, and 25 ms window length. An examplary log Mel-spectrogram is depicted in Fig. 3.1. While for the extraction of MFCCs with $\Delta\&\Delta\Delta$ this spectro-temporal representation is processed spectrally with a DCT and temporally with slope-filters, GBFB features are extracted with 2D-Gabor filters that perform a simultaneous spectral and temporal processing. Fig. 3.2 depicts the relation of the spectro-temporal 2D-Gabor filters and the effective MFCC-DD spectro-temporal patterns. The outer product of a DCT base function and a Delta base function gives the effective spectro-temporal pattern that the corresponding MFCC-DD dimension encodes. The GBFB feature extraction is illustrated in Fig. 3.3. First, spectro-temporal patterns are extracted by 2D-convolving the 2D-Gabor filter functions with the log Mel-spectrogram. A subsequent selection of representative channels by critically sampling the filtered log Mel-spectrograms limits the systematical correlation of the feature dimensions. Each 2D-Gabor filter extracts patterns of a pair of a spectral and a temporal modulation frequency. These features were shown to improve the robustness of a MFCC baseline system when fed directly into an GMM/HMM recognizer (Schädler et al., 2012a). The range of modulation frequencies covered is about 6 to 25 Hz and 0.03 to 0.25 cycles/Mel-band. Some properties of the GBFB features are compared with those of MFCC

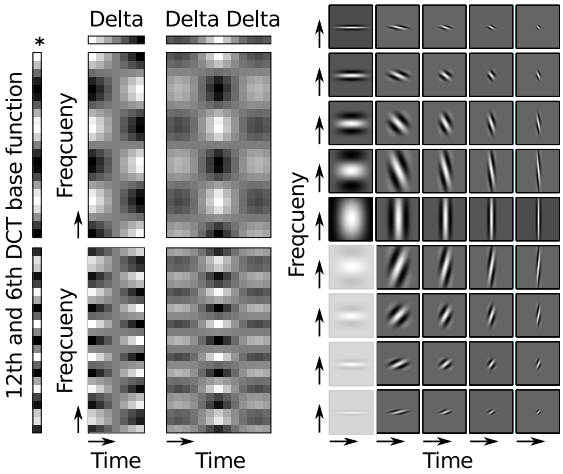

Figure 3.2: *Left panel: Effective spectro-temporal patterns of combined traditional spectral DCT and temporal Δ&ΔΔ processing. Right panel: The 41 2D-Gabor filters that are used for feature extraction with the Gabor filter bank. The patterns are scaled and their real spectral extension is the same as of the MFCC-DD patterns in the left panel.*

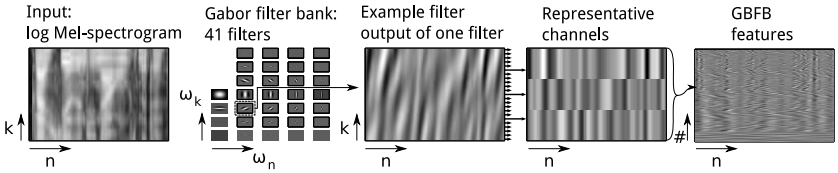

Figure 3.3: *Illustration of the Gabor filter bank feature extraction. n: temporal index; k: spectral index; ω: modulation frequencies. The input log Mel-spectrogram is filtered with each of the 41 filters of the Gabor filter bank. Representative channels of the filter outputs are selected and concatenated. The 311-dimensional output is used as feature vector.*

Table 3.1: *Properties of MFCCs and GBFBs compared*

Feature	Spectral	Temporal	Separable	Dimension
MFCC-DD	DCT	$\Delta\&\Delta\Delta$	yes	39
GBFB	Gabor	Gabor	no	311

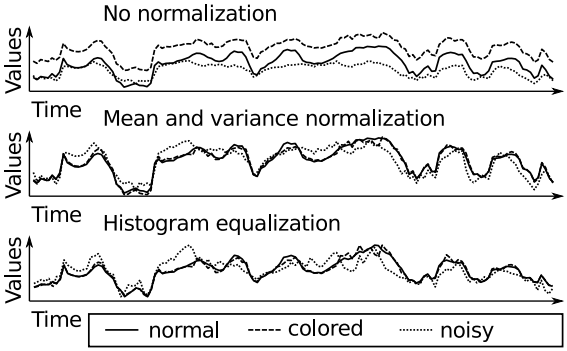

Figure 3.4: *Illustration of mean and variance normalization and histogram equalization of the first MFCC values for a speech signal in different acoustic contexts.*

features in Tab. 3.1. The MFCC-DD processing can be described by separate spectral and temporal operations, while the GBFB processing cannot.

3.2.2 Normalization of feature value statistics

It has been shown that the robustness of an ASR system with MFCC features can be increased by removing the mean value and normalizing the variance of each feature dimension (Viikki and Laurila, 1998). This processing is called mean and variance normalization (MVN) and normalizes the first and the second moments of the feature value distributions. An extension to MVN is mapping the feature values to a specific reference distribution (De La Torre et al., 2005). This processing is called histogram equalization (HEQ) and normalizes all moments of the feature value distributions. The effect of MVN and HEQ on the first (not zeroth) MFCC is illustrated in Fig. 3.4. A spectral coloring

(eg. preemphasis) of a speech signal leads to a systematic changes in the log Mel-spectrogram and consequently to a change of the derived features (cf. offset/mean value in Fig. 3.4). Likewise, additive noise or reverberation result in a reduction of the dynamic range by filling up the "valleys" of the log Mel-spectrogram, which may be interpreted as a reduction of modulation depth (cf. scale/variance in Fig. 3.4 *noisy*). Applying MVN/HEQ to MFCC/GBFB features counteracts the influence of the most common sources of variability in noisy speech by normalizing the modulation depth, because the feature values scale linearly with it. The recognition performance of GBFB and MFCC features is evaluated with and without MVN and HEQ.

3.2.3 Recognition experiment and baseline

The effect of the different front-ends on the robustness of an ASR system is evaluated within the Aurora 2 framework (Pearce and Hirsch, 2000). The task is the recognition of English connected digits which are contaminated with eight different everyday background noises from 20 dB to −5 dB. The framework provides speech data for training and testing as well as a GMM/HMM classifier and trainings rules. A reference setup defines whole-word left-to-right HMMs with 16 states, 3 mixtures per state, and without skips over states. The back-end is not modified and used with the same parameters as in the reference. Two different training conditions exist. For *clean* training only utterances *without* added noise are used, while for *multi* training utterances *with and without* added noise are used. Although only four noise types that occur in the testing data are also included in multi training data, it allows the recognizer to learn the reliability of feature patterns in noise. As reference features the first 13 MFCCs with first and second order discrete temporal derivative ($\Delta\&\Delta\Delta$) are used, resulting in 39-dimensional MFCC-DD features. Additionally the baseline results for ETSI MFCC (ETSI, 2003, Standard 201 108 v1.1.3) and ETSI Advanced Front-End (AFE) (ETSI, 2007, Standard 202 050 v1.1.5) features are reported. The word recognition accuracies are compared at signal-to-noise ratios (SNR) from 20 to −5 dB.

3.2.4 Spectral and temporal contribution

With the aim of evaluating the effect of normalizing only spectral, only temporal, or spectro-temporal patterns, the separability of spectral and

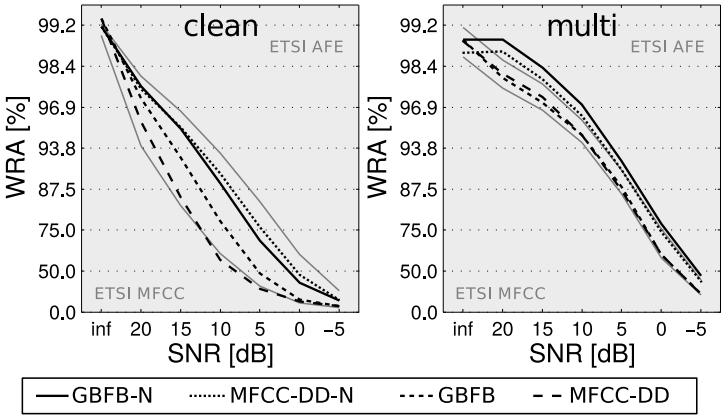

Figure 3.5: *Word recognition accuracies for GBFB and MFCC-DD feature with and without mean and variance normalization (N) at different test signal to noise ratios and training styles.*

the temporal processing with MFCC-DD features is exploited The normalization (N) is applied at the following stages of MFCC-DD feature calculation: MFCC-N-DD, MFCC-DD-N, DD-N-MFCC. With MFCC-N-DD features, spectral patterns are integrated by the DCT before normalization. With DD-N-MFCC features, short term temporal patterns are integrated by the $\Delta\&\Delta\Delta$ processing before normalization. And with MFCC-DD-N features, spectral and short term temporal patterns are integrated before normalization. The recognition performance of the differently normalized features is evaluated.

3.3 RESULTS AND DISCUSSION

3.3.1 Normalized GBFB features

Average word recognition accuracies (WRA) for GBFB and MFCC features with and without MVN are reported in Fig. 3.5. With *clean* condition training MVN dramatically improves the robustness of MFCCs by $5-7$ dB over a wide range of WRAs (50% to 95%). The improvements for GBFBs with $2-3$ dB are smaller, but they perform about 3 dB better without MVN. Thus, MFCCs perform about 1 dB better than GBFBs at

low SNRs, but cannot improve the highly optimized ETSI AFE baseline. However, GBFB features outperform all features when testing on clean data. In terms of average relative improvement over SNRs from 20 dB to 0 dB, MFCCs with MVN improve the WRA of the ETSI MFCC baseline by 58% on, while GBFBs with MVN improve the baseline by 54%. With *multi* condition training MVN improves the performance of MFCCs almost independently of the SNR by about $2-3$ dB. For GBFB features the improvements are with 2.5 dB at low SNRs and up to 6 dB at high SNRs more pronounced. In terms of average relative improvement over SNRs from 20 dB to 0 dB, MFCCs with MVN improve the WRA of the ETSI MFCC baseline by 37%, while GBFBs with MVN improve the baseline by 48%. GBFB features outperform all other features, including ETSI AFE, in every noisy testing condition. The the improvements with HEQ were found to be similar to the improvements with MVN within a range of ± 1 dB and are therefore omitted. The very high recognition scores for clean testing data with clean condition training, as well as for high SNRs with multi condition training (which contains speech data at ∞, 20, 15, 10, and 5 dB SNR) indicate a certain sensitivity of GBFB features to mismatched SNR conditions. Possibly, the 311-dimensional GBFB features encode more precise information about the speech signal than the 39-dimensional MFCC features which results in a higher sensitivity to the SNR. This finding puts the one-model-for-all-SNRs approach into question, as speech at 0 dB SNR and speech at 20 dB SNR have quite different characteristics. If the hypothesis holds, than GBFB features with MVN should perform even better in context-dependent models, which should be evaluated in future experiments.

3.3.2 Spectral vs. temporal normalization

Average word recognition accuracies (WRA) for MFCC features with and without MVN of spectral, temporal, and spectro-temporal patterns are depicted in Fig. 3.6.

Normalizing the output after the temporal processing and before the spectral processing results in worse performance than without normalization, with an exception at very low SNRs with *multi* condition training. The MVN effectively normalizes all Mel-bands to have the same energy and the same modulation depth which seems to accompanied by a loss of information that is relevant for robust ASR. Normalizing the output after the spectral processing and before the temporal processing results

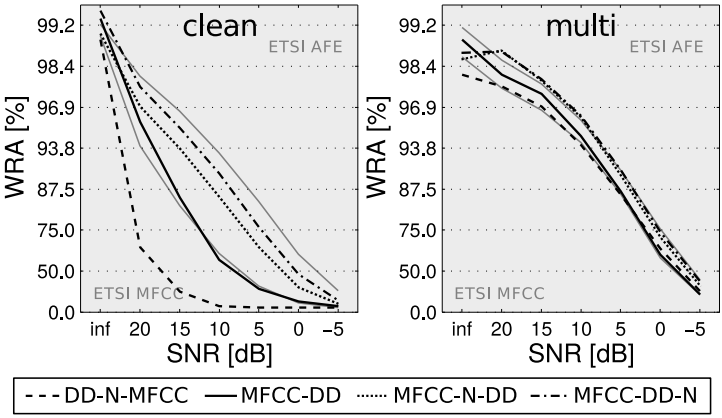

Figure 3.6: *Word recognition accuracies for MFCC-DD features with and without mean and variance normalization of spectral (MFCC-N-DD), temporal (DD-N-MFCC), and spectro-temporal (MFCC-DD-N) patterns at different signal to noise ratios for clean and multi style training.*

in important improvements, but the best performance is achieved by normalizing after the spectral and temporal processing. This indicates that spectro-temporal patterns are best extracted from an unprocessed spectro-temporal representation and normalization is best performed after spectral and temporal integration.

3.4 CONCLUSIONS

The most important findings of this work can be summarized as follows:

— Normalization increases the robustness of physiologically motivated spectro-temporal Gabor filter bank features by 2.5 – 5.0 dB SNR on a digit recognition task, outperforming ETSI AFE features with multi-style training.

— Normalization of separable spectro-temporal patterns was found to be best applied after spectral and temporal integration.

— Normalized Gabor filter bank features seem work well in matched signal to noise ratio conditions, which should be further investigated with SNR-dependent models.

ACKNOWLEDGMENTS

This work is funded by DFG SFB/TRR 31 “The active auditory system”.

4 High-dimensional spectro-temporal auditory features for robust medium-size vocabulary speech recognition systems

ABSTRACT

The robustness of high-dimensional auditory spectro-temporal features is evaluated on a medium-size vocabulary speech recognition task. For this, the data of second track of the second CHiME challenge is employed, which offers a 5k-word speech recognition task in a realistic noisy environment. The auditory features, called Gabor filter bank features, are compared to a Mel Frequency Cepstral Coefficient (MFCC) baseline. However, for a fair comparison a set of parameters has to be tuned for each feature type. An algorithm that finds the needed parameter values for a fair comparison of different feature sets is proposed. We find a 1.54 percentage point improvement in word recognition accuracy by using GBFB features. While performing better than the baseline features, the use of GBFB features does not entail increased computational requirements.

4.1 INTRODUCTION

Automatic speech recognition (ASR) systems still perform worse than human listeners in almost any kind of speech recognition task. This is especially true in acoustically adverse conditions with reverberation and noise sound at signal-to-noise ratios (SNR) below 0 dB. There are ongoing efforts to narrow the gap between human and machine speech recognition performance under realistic acoustic conditions, like e.g., in the second CHiME Challenge (Vincent et al., 2013a). Besides purely technical approaches to improve the system's performance at very low SNRs, attempts exist to integrate knowledge about the human auditory system—which performs reasonably well—into ASR systems.

One successful approach to increase the robustness of an ASR system by integrating auditory processing turned out to be the extraction of

normalized spectro-temporal patterns. These patterns are extracted with a filter bank of physiologically motivated 2D Gabor filters (Schädler et al., 2012a) and have been found to improve the robustness of a standard ASR system on a digit in noise recognition task by 2.5 dB to 5.0 dB. (Schädler and Kollmeier, 2012b). Recently, this finding was confirmed and extended to letters in noise (Moritz et al., 2013).

The aim of this study is to evaluate the robustness of GBFB features on a speaker independent medium-size vocabulary task with 5000 words using a bigram language model under difficult acoustic conditions. Unlike many small-size vocabulary tasks that allow the use of whole-word acoustic models, medium-size vocabulary tasks require acoustic models for smaller fragments of speech, such as triphones. The parameters of triphone acoustic models are usually tied with threshold-governed clustering techniques in the training process because sufficient training data is not available for all models. Because a good state-tying threshold depends on many factors—the feature type and dimension being two of them—it has to be found for each new feature or training data set in order to be able to perform a fair comparison. Further, medium-size vocabulary speech recognition tasks require non-binary language models that have to be weighted with a *grammar scale factor* that also has to be found for each new feature or training data set. Finally, an exhaustive search of the whole state-space for decoding with such a complex language model is infeasible and usually avoided by pruning unlikely paths. Hence, a search beam pruning threshold, which determines the search depth and modulates the needed computation time for recognition, also needs to be found for each new feature or training data set. Therefore, it is legitimate to use computation time as a soft constraint, and for a fair comparison computation time should be similar across the evaluated systems.

The extraction of Gabor filter bank features is explained in Sec. 4.2.1. The recognition experiment and the baseline features are presented in Sec. 4.2.2. A solution for the problem of finding the state tying thresholds, the grammar scale factors, and the search beam pruning thresholds is proposed in Sec. 4.2.3. Results are presented, compared, and discussed in Sec. 4.3, while Sec. 4.4 concludes the study.

4.2 METHODS

4.2.1 Gabor filter bank features

The Gabor filter bank (GBFB) feature extraction, which is motivated by auditory processing, encodes spectro-temporal modulation patterns in high-dimensional feature vectors (Schädler et al., 2012a). This feature extraction scheme improves the robustness of a standard ASR system on a digit in noise recognition task by $2.5-5.0$ dB compared to a system with traditional Mel Frequency Cepstral Coefficient (MFCC) features (Schädler and Kollmeier, 2012b). GBFB features are extracted by applying a set of 2D Gabor filters to a spectro-temporal representation of a signal. A log Mel-spectrogram is employed as the spectro-temporal representation because it incorporates several properties of the auditory system (i.e., non-linear frequency scaling and compression of amplitude values) and is widely used in ASR. In contrast to the GBFB reference implementation which uses a log Mel-spectrogram with 23 Mel-bands between 64 Hz and 4 kHz, 10 ms window shift, and 25 ms window length, we extend the frequency range to 8 kHz and increase the number of Mel-bands to 31. This results in the lower 23 Mel-bands to cover the frequency range from 64 Hz to ≈ 4 kHz, like in the original set-up. The shapes of the 41 GBFB filters that are used for feature extraction are shown in Fig. 4.1. The GBFB feature extraction steps are illustrated in Fig. 4.2.

First, spectro-temporal patterns are extracted by 2D-convolving 2D Gabor filter functions with the log Mel-spectrogram. Each filter extracts patterns of a different pair of a spectral and a temporal modulation frequency, where the center modulation frequencies range from 6 to 25 Hz, and from 0.03 to 0.25 $\frac{\text{cycles}}{\text{Mel-band}}$ (cf. Fig. 4.1). A subsequent selection of representative channels by critically sampling the filtered log Mel-spectrograms limits the systematic correlation of the feature dimensions and reduces the number of dimensions of the feature vector from $31 \times 41 = 1271$ to 455 (Schädler et al., 2012a). Mean and variance of each of the 455 feature dimensions are normalized on an utterance basis. In the following these features are referred to as GBFB features or just GBFBs.

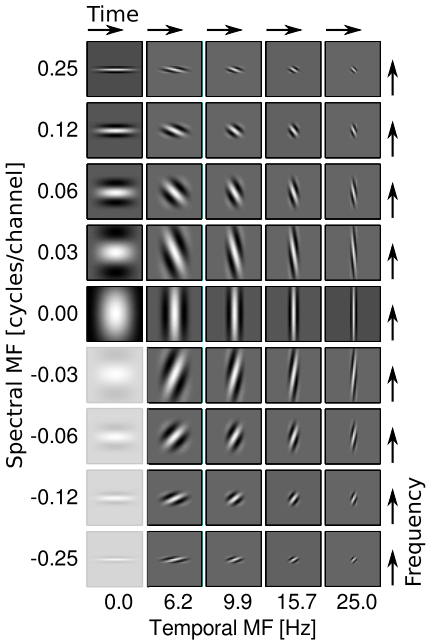

Figure 4.1: *2D filter functions of the* 41 *Gabor filters used for the filter bank, arranged by temporal modulation frequencies (MFs)* ω_n *and spectral MFs* ω_k*. Within each tile, the horizontal and vertical axis represent time and frequency, respectively. Duplicate filters that occur due to symmetries in the GBFB are greyed out and not used.*

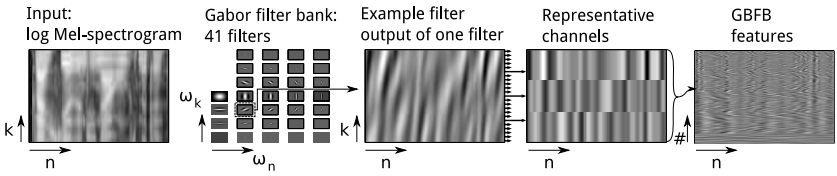

Figure 4.2: *Illustration of the Gabor filter bank feature extraction. n: temporal index; k: spectral index;* ω*: modulation frequencies. The input log Mel-spectrogram is filtered with each of the* 41 *filters of the Gabor filter bank. Representative channels of each of the filter outputs are selected and concatenated. The* 455*-dimensional output is used as feature vector.*

4.2.2 Recognition experiment and baseline

The task that is employed to evaluate the robustness of GBFB features is the recognition of utterances being spoken in a noisy living room from recordings made using a binaural mannikin. We employ the isolated training and development data sets from the second track of the second CHiME challenge (Vincent et al., 2013a), which are based on the Wall Street Journal (WSJ0) corpus, a corpus of read speech. The utterances of these data sets are filtered with a fixed binaural room impulse response of a living room corresponding to a frontal position at a distance of 2 m. Subsequently, the utterances are mixed with noise samples recorded in that living room at SNRs from −6 dB to 9 dB.

The training and testing scripts provided in the CHiME challenge are based on HTK (Young et al., 2006) and on HTK recipes from Vertanen (2006) for WSJ. We developed an own set of training and testing scripts using HTK and the training protocol described by Vertanen (2006), and adapted the testing scripts in order tune the needed parameters in an automatic fashion. The speech recognizer employs 3-state left-to-right triphone acoustic models with up to 8 gaussian mixtures assuming diagonal covariance matrices, a 3-state 16-mixture background model with skip and back transitions, and a short pause model tied to the center state of the background model. The CMU Pronouncing Dictionary (Carnegie Mellon University, 2007) version 7a is employed to generate initial monophone labels. The questions that are needed for the tree-based state tying are taken from Vertanen (2006). For recognition, the same 5k-word bigram language model as in the second track of the second CHiME challenge is used.

The system is trained on the isolated training set, which contains 7138 utterances from 83 speakers mixed with background noise at random SNRs from −6 dB to 9 dB. The performance is evaluated on the development set, which contains 409 utterances from 10 other speakers, at the following SNRs: −6, −3, 0, 3, 6, and 9 dB. Mel Frequency Cepstral Coefficients (MFCCs) serve as baseline features. The baseline features are extracted by first calculating 13 (including 0th) MFCCs from a log Mel-spectrogram with 26 Mel-Bands from 64 Hz to 8 kHz. To these, a slope filter with a total width of 5 frames is applied in order to calculate delta and acceleration coefficients. The MFCCs with delta and acceleration coefficients form a 39-dimensional feature vector of which for each dimension mean and

variance are normalized on an utterance basis. In the following, these features are referred to as MFCC features or just MFCCs.

4.2.3 Parameter search

For the recognition experiment described in Sec. 4.2.2, values for the following parameters have to be found: A *state tying threshold* that determines the total number of tied states. A *grammar scale factor* which determines the weight of the language model. And a *search beam pruning threshold* that limits the depth of the search of the state sequence that best explains the observed feature vectors. These values depend—not exclusively—on: the feature type and dimension, the training data type and amount, the recognition task, and the language model. For a fair comparison of systems with different features, an algorithm to find these values is required.

In HTK, the threshold for the tree-based state clustering, the language model values that are scaled by the grammar scale factor, and the search-beam pruning thresholds, are all likelihoods. The maximum log-likelihood of a separable n-dimensional normal distribution with the same variance in each dimension is proportional to n. This relationship between dimensionality and log-likelihood predicts a factor of about 10 for log-likelihoods with 455-dimensional GBFB features and 39-dimensional MFCC features. For MFCCs a good value for the state-tying threshold can be expected between 0 and 1000 (Young et al., 2006; Vertanen, 2006). Applying the estimated factor of 10, for GBFB features, we expect a good value between 0 and 10000. Hence, for MFCCs, systems are trained with state tying threshold of 0, 200, 300, 400, 500, 600, 800, and 1000, and for GBFBs, systems are trained with state tying threshold of 0, 2000, 3000, 4000, 5000, 6000, 8000, and 10000.

For evaluation of the trained systems, suitable values for the grammar scale factor and the beam pruning threshold that provide a good recognition accuracy in feasible computation time still have to be found. As the grammar scale factor modulates the combined likelihood of the acoustic and the language model, a good value for the *pruning* threshold depends on the value for the grammar *scale*. To find a good pair of values (*scale* and *pruning*), 20 randomly chosen utterances of the development at 0 dB SNR are used as a tuning set. Only utterances with a SNR of 0 dB are selected because the used background noises are highly unstationary and cover a wide range of short-term SNRs. Available computational

resources demand the low number of utterances in the tuning set that is used across features and conditions.

The values for *scale* and *pruning* are determined as follows: For the first iteration ($i = 1$), values considered for scale are $2^{0,1,\ldots,8}$, and for pruning $2^{2,3,\ldots,12}$. Time-limited recognition processes are started with all combinations of the considered values. The time limit for recognition was set to 40 minutes for the 20 utterances in the tuning set, which corresponds to a lower limit for recognition speed of 0.13 words per second. The value pairs that result in the best recognition performance determine the region of interest for the next iteration, where only value pairs that result in a recognition of all 20 utterances within the given time are considered valid. Let s and p be the set of scale and pruning values of the best performing combinations, respectively. Let ls and lp be the binary logarithm of s and p, respectively. For the next iteration, the search grid is refined by halving the step width sw of the exponent (1 in the first iteration) of the actual iteration. For scale, the following values are considered for the next iteration: $2^{\min(ls)-sw,\ldots,\max(ls)+sw}$; and for pruning: $2^{\min(lp)-sw,\ldots,\max(lp)+sw}$. In this study, four iterations were performed. Of the valid scale/pruning value combinations with the highest recognition performance, the one that required the least time to complete the recognition is chosen.

Grammar scale factor and pruning thresholds are automatically determined for each trained system using the described algorithm, and a security margin of 20% is added to the determined pruning thresholds. The results are presented and compared in Sec. 4.3.

4.3 RESULTS AND DISCUSSION

The automatically determined parameters and the recognition results of the medium-size vocabulary speech in noise recognition task for ASR systems with MFCC and GBFB features with different state tying thresholds are reported in Table 4.1. The system with the 39-dimensional MFCC features covers the range from about 1.2 k to 9.5 k tied states with state tying thresholds between 0 and 1000. With GBFB features, state tying thresholds between 0 and 10000 cover roughly the same range of tied states. This result indicated the applicability of the approximated relationship of feature dimensionality and log-likelihood in Sec. 4.2.3. The recognition accuracies depending on the number of tied states with MFCC

Table 4.1: *Automatically determined parameters and recognition results of ASR systems using MFCC and GBFB features with different state tying thresholds. In the column 'Speed', the recognition speed in words per second, and in the column 'Tuning Accuracy', the word recognition accuracy of the utterances in the tuning set is reported. Lines corresponding to the systems with the best mean recognition performance for MFCCs and for GBFBs stand out in bold type.*

	Tying thr.	Tied states	Scale factor	Prun. thr.	Speed $[\frac{\text{w}}{\text{s}}]$	Tuning acc. [%]	Word recognition accuracy [%] −6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	mean
MFCC	0	9559	14.7	197.4	0.16	43.99	28.19	34.75	41.41	45.74	51.62	55.51	42.87
	200	5756	12.3	166.0	0.16	51.27	30.76	37.72	44.76	49.42	55.66	59.02	46.22
	300	3598	11.3	166.0	0.13	49.05	31.77	37.07	45.29	49.56	56.50	60.26	46.74
	400	**2712**	**11.3**	**152.2**	**0.22**	**51.27**	**31.16**	**38.53**	**44.95**	**50.60**	**56.79**	**60.92**	**47.16**
	500	1991	14.7	181.0	0.21	50.63	31.86	37.20	44.95	49.12	54.45	58.03	45.94
	600	1827	13.5	181.0	0.15	51.90	31.55	37.66	44.96	49.67	54.68	58.98	46.25
	800	1451	13.5	166.0	0.21	47.78	31.51	37.09	44.34	48.83	54.90	57.57	45.71
	1000	1206	12.3	166.0	0.16	46.84	29.72	35.79	42.93	48.09	53.43	55.97	44.32
GBFB	0	9543	64.0	939.0	0.19	42.72	27.45	32.00	38.39	41.54	47.44	50.97	39.63
	2000	5877	69.8	1116.7	0.13	49.37	32.11	37.22	43.76	48.99	53.99	58.36	45.74
	3000	3486	64.0	861.1	0.29	52.85	32.88	37.69	46.50	50.74	56.90	61.38	47.68
	4000	**2486**	**69.8**	**939.0**	**0.26**	**53.16**	**34.18**	**39.36**	**47.43**	**52.00**	**57.43**	**61.81**	**48.70**
	5000	1981	64.0	939.0	0.18	53.80	34.42	39.84	47.53	51.84	57.41	61.14	48.70
	6000	1634	64.0	789.6	0.47	49.05	32.94	38.87	46.29	50.04	56.73	60.32	47.53
	8000	1264	90.5	1116.7	0.24	49.68	33.53	38.83	45.55	49.74	56.01	59.54	47.20
	10000	1054	64.0	861.1	0.30	47.47	32.19	38.74	44.86	49.42	54.51	59.65	46.56

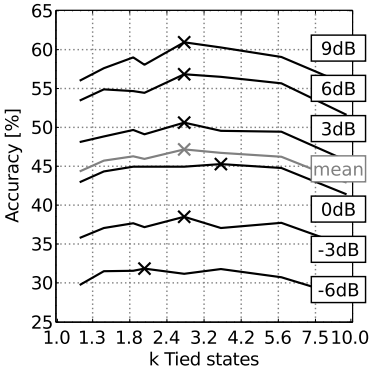

Figure 4.3: *Recognition accuracies depending on the number of tied states with MFCC features at different test SNRs.*

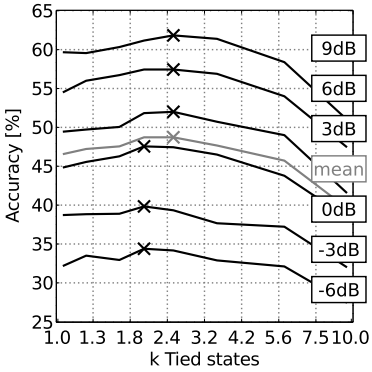

Figure 4.4: *Recognition accuracies depending on the number of tied states with GBFB features at different test SNRs.*

features are plotted in Fig. 4.3, and with GBFB features in Fig. 4.4. The best performing system is marked with a cross for each SNR.

The optimal number of tied states for this set-up seems to be somewhere around 2.4 k tied states, which indicates that it could suffice to compare different systems with a fixed number of tied states. Lower and higher numbers result in decreased recognition performance. This effect is more pronounced for high than for low test SNRs, and for GBFB features than for MFCC features. For MFCCs, the optimal number of tied states does not depend on the noise energy in the tested utterances, while for GBFBs there might be a slight trend towards less tied states at low SNRs. On average over all SNRs, systems with less than about 4 k tied states perform better with GBFB than with MFCC feature, with more tied states the situation is reversed. For all SNRs, the best performing system uses GBFB features. On average over all SNRs, the best value for the state tying threshold is 400 for MFCCs, and 4000 for GBFBs (cf. line in bold type in Table 4.1). Comparing these systems, GBFB features improve the MFCC baseline by 1.54 percentage points on average. This corresponds to an improvement of about 0.74 dB on average over all SNRs, which is less than the improvement of at least 2 dB that is observed on small-size vocabulary speech recognition tasks.

On one hand, this could be due to the small amount of speech material for the high number of relevant speech fragments that have to be modelled, resulting in less data per relevant model compared to a less complex task. On the other hand, the recognition of up to 5000 different words is a more complex task than to tell apart digits or letters, and some "bottleneck" of this highly nonlinear system could modulate the effect of the features. Despite the high dimensionality of the GBFB features, the difference in computation time that is needed for recognition with GBFBs and with MFCCs is negligible. Recognition speed depends strongly on the chosen search beam pruning threshold and the "clarity" of the observations; e.g., noisy utterances usually take more time to be recognized on the same hardware. The total of 6779 words was recognized at a rate of 6.0 words per minute at 9 dB SNR and at a rate of 3.2 words per minute at −6 dB SNR with MFCC features and 2712 tied states. With GBFB features and 2486 tied states, 7.6 words were recognized per minute at 9 dB SNR and 4.1 at −6 dB SNR. This can be explained either with the evaluation of the observation probability being much less computationally expensive than the other decoding steps, or with GBFB features offering clearer observations that result in better pruning decisions and thus allow lower beam-pruning thresholds.

4.4 CONCLUSIONS

The most important findings of this work can be summarized as follows:

— High-dimensional auditory Gabor filter bank (GBFB) features improve the robustness of an ASR system on a 5k vocabulary word recognition task under realistic adverse acoustic conditions compared to MFCC features.

— For a fair comparison of different feature types on a medium-size vocabulary speech recognition task, an algorithm for tuning a set of parameters is needed. Such algorithm has been proposed.

— ASR with high-dimensional GBFB features is as fast as with MFCC features.

ACKNOWLEDGMENTS

This work was funded by DFG SFB/TRR 31 "The active auditory system".

5 Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition

ABSTRACT

To test if simultaneous spectral and temporal processing is required to extract robust features for automatic speech recognition (ASR), the robust spectro-temporal two-dimensional-Gabor filter bank (GBFB) front-end from Schädler, Meyer, and Kollmeier [J. Acoust. Soc. Am. 131, 4134–4151 (2012)] was de-composed into a spectral one-dimensional-Gabor filter bank and a temporal one-dimensional Gabor filter bank. A feature set that is extracted with these separate spectral and temporal modulation filter banks was introduced, the separate Gabor filter bank (SGBFB) features, and evaluated on the CHiME (Computational Hearing in Multisource Environments) keywords-in-noise recognition task. From the perspective of robust ASR, the results showed that spectral and temporal processing can be performed independently and are not required to interact with each other. Using SGBFB features permitted the signal-to-noise ratio (SNR) to be lowered by 1.2 dB while still performing as well as the GBFB-based reference system, which corresponds to a relative improvement of the word error rate by 12.8%. Additionally, the real time factor of the spectro-temporal processing could be reduced by more than an order of magnitude. Compared to human listeners, the SNR needed to be 13 dB higher when using Mel-frequency cepstral coefficient features, 11 dB higher when using GBFB features, and 9 dB higher when using SGBFB features to achieve the same recognition performance.

This chapter is a reformatted reprint. The original article can be found at http://dx.doi.org/10.1121/1.4916618. Reproduced with permission from "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition", M. R. Schädler and B. Kollmeier, J. Acoust. Soc. Am. Vol. 137, pp. 2047–2059. Copyright 2015, Acoustical Society of America.

5.1 INTRODUCTION

After years of investigation on robust automatic speech recognition (ASR), human listeners still outperform ASR systems in realistic acoustic environments (Lippmann, 1997; Meyer et al., 2011b; Barker et al., 2013). Inspired by the ability of the human auditory system to decode speech signals in the most difficult acoustic conditions, many principles of auditory signal processing were integrated into ASR systems in attempts to improve their recognition performance. These approaches usually targeted the feature extraction stage (front-end), where the more tangible peripheral auditory processes can be mapped to signal processing algorithms and which is more specific to auditory processes than the recognition stage (back-end). The current study aimed to improve the front-end by extracting spectro-temporal modulation features with independent spectral and temporal processing instead of joint spectro-temporal processing.

Many of the speech representations (or features) used in ASR systems stem from spectro-temporal representations of sound that already incorporate basic auditory principles, such as the log Mel-spectrogram (LMSpec). The LMSpec is a spectrogram with a logarithmic amplitude and a Mel frequency scaling. It considers very basic auditory principles of the human auditory system, such as the resolution across frequencies and logarithmic perception of intensity. However, these static spectro-temporal representations themselves are not well suited as robust speech features because environmental changes, such as additive noise and reverberation, strongly affect them. The characteristics of the inherently dynamic speech signals are better represented in changes that occur in the spectro-temporal representations across frequencies and over time; this is why many robust features are extracted by encoding spectral or temporal changes. An example for spectral processing is the still widely used Mel-frequency cepstral coefficients (MFCCs), which perform a discrete cosine transform in the spectral dimension of a LMSpec (Davis and Mermelstein, 1980). An example for temporal processing is the calculation of discrete temporal first and second order derivatives, called deltas and double deltas, which are usually used to encode the dynamics of MFCC and other features. Many other, differently motivated spectral and temporal processing schemes were combined with the goal of improving the robustness of ASR systems (e.g., Hermansky, 1990; Hermansky et al., 1992; Hermansky and Sharma, 1999; Nadeu et al., 2001; Hermansky

and Fousek, 2005; Moritz et al., 2011) but without relating the spectral to the temporal processing nor vice versa.

In approaches to join spectral and temporal modulation processing, and thus allowing for higher order dependencies between both, Kleinschmidt (2002b) and Kleinschmidt and Gelbart (2002a) found that the physiologically motivated (Qiu et al., 2003) two-dimensional (2D) spectro-temporal Gabor filters were good candidates. Aside from their use in ASR systems, a number of studies suggested the use of 2D Gabor filters to extract spectro-temporal features for acoustic signal and speech analysis, (e.g., Chi et al., 2005; Mesgarani et al., 2006; Ezzat et al., 2007b). Because in early approaches to extract features with 2D Gabor filters the filter parameters were determined in a data driven way, and as a consequence some feature dimensions were highly correlated, Meyer and Kollmeier (2011a) mapped these Gabor features to an intermediate phoneme probability layer by means of a tandem setup to use them with standard Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) based recognition back-ends. Recently, in an approach to structure the 2D Gabor filter parameter space and gain a set of universal 2D Gabor filters for robust speech recognition, the 2D Gabor filter bank (GBFB) features were introduced and shown to improve the robustness of ASR systems when they are used directly with standard GMM/HMM back-ends by Schädler et al. (2012a) and Moritz et al. (2013). The 2D spectro-temporal filters of the GBFB, which were used to extract robust speech features by 2D-convolving each of them with a LMSpec, are depicted in Fig. 5.1 and cover a range of spectral and temporal modulation frequencies that were found to be beneficial for robust ASR. The extraction of GBFB features is explained in detail in Sec. 5.2.2. Meyer and Kollmeier (2011a) attributed the improvements in robustness to a locally increased SNR due to the higher sensitivity to speech patterns of the more complex spectro-temporal patterns, most notably to the ability of discriminating upward and downward spectro-temporal patterns (cf. off-axis filters in Fig. 5.1). Schröder et al. (2013) found that using GBFB features can improve the recognition performance in a speech-unrelated acoustic event detection task; this confirms the universality of the GBFB filter set for acoustic recognition tasks. However, a model of joint spectro-temporal processing does not allow changes to the spectral processing without having an effect on the temporal processing and vice versa; this would imply that all models of separate spectral and temporal processing are

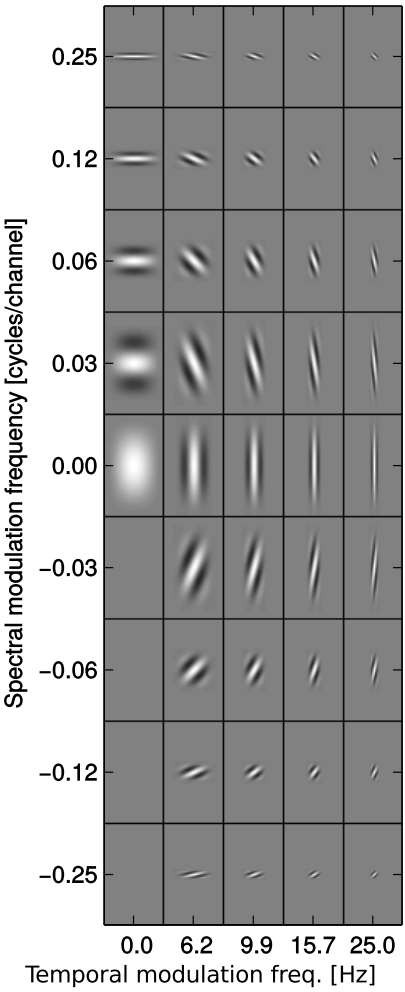

Figure 5.1: *Taken from Schädler et al. (2012a). Filter shapes of the 2D Gabor filter bank (GBFB) filters. Each tile represents the filter function of a spectro-temporal 2D Gabor filter, where the horizontal axis within each tile is the temporal one and the vertical axis is the spectral one. The 2D filter functions are sorted by their spectral and temporal center modulation frequencies. To extract GBFB features, a LMSpec of speech is filtered by means of a 2D convolution with these filters. While the filters on the axis (0 Hz or 0 cycles/channel) are purely spectral or purely temporal filters and can be separated into a real-valued spectral 1D filter and a real-valued temporal 1D filter, the off-axis filters are inseparable spectro-temporal filters.*

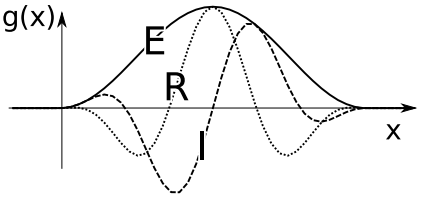

Figure 5.2: *Absolute (E), real (R), and imaginary (I) part of a complex-valued filter function of a 1D Gabor filter with* 3.5 *half-waves under the envelope. Each part is a real-valued function and can be used to filter a signal where R and I are band-pass filters with the same transfer function and only differ in the phase, while E describes a low-pass filter.*

insufficient. It is unknown to what extent spectral and temporal processing in the auditory system of mammals interact with each other (Depireux et al., 2001; Qiu et al., 2003). Further, the more complex 2D filtering process results in considerably higher computational costs for the feature extraction. If spectral and temporal processing were independent processes, the mentioned limitations would not apply.

In this study, it was investigated whether the improvements in robustness gained with the structured, spectro-temporal GBFB approach require the complex joint 2D spectro-temporal processing or if a separate spectral and temporal processing with two 1D GBFB can be used to extract features that perform similarly or better. The basic idea was to replace the inseparable up- and downward 2D patterns of the GBFB with separable patterns and then perform the spectral and the temporal filtering separately with 1D Gabor filters. A 1D Gabor filter is depicted in Fig. 5.2 and the relation of 1D-spectral and 1D-temporal Gabor filters to the inseparable up- and downward 2D-spectro-temporal Gabor filters is illustrated in Fig. 5.3.

In Fig. 5.3, it can be observed that the addition (A)/subtraction (S) of an inseparable 2D spectro-temporal downward (D) filter to/from its corresponding upward (U) filter is identical to the separable filter RR/II, which in turn can be described by a separate spectral and temporal filtering process with the real (R) or imaginary (I) part of 1D Gabor functions. The relation between a pair of a spectral and a temporal 1D filter, and the corresponding 2D filter is the outer product and is

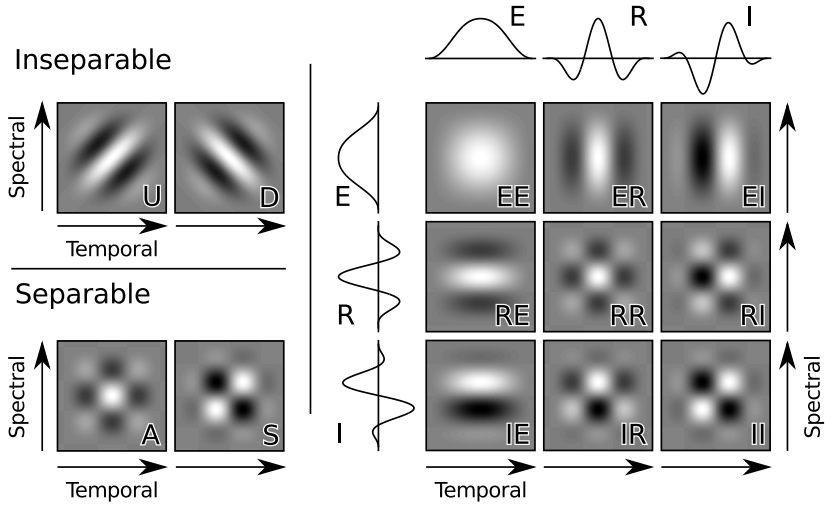

Figure 5.3: *Inseparable and separable 2D spectro-temporal Gabor filters and their relation to separate 1D spectral and 1D temporal Gabor filters. Each tile represents the filter function of a 2D spectro-temporal filter with the horizontal axis within each tile being the temporal and the vertical axis being the spectral one. Left panel: The 2D upward (U) and downward (D) filters are not separable, while their sum A (U + D) and difference S (U - D) are; right panel: Effective 2D filter shapes when applying subsequent spectral and temporal 1D filters using different parts of 1D Gabor filters (E, envelope; R, real part; I, imaginary part). The amplitude of the 2D filters is encoded in gray scale, where white means high amplitude and black low amplitude.*

explained later in more detail. The combination of spectral and temporal filters with different phases, which were determined by the use of the real (R) or imaginary (I) part, but identical center modulation frequencies resulted in different effective spectro-temporal filter patterns (cf. RR, RI, IR, and II in Fig. 5.3). Hence each inseparable 2D filter in Fig. 5.1 could have been replaced with different separable 2D filters that have the same absolute spectral and temporal center modulation frequencies as the inseparable 2D filter.

Instead of only replacing the inseparable 2D filters, the whole 2D GBFB was replaced by two separate 1D GBFB: A spectral one and a temporal one. For these, the positive spectral and temporal center modulation frequencies were taken from the 2D GBFB. The phase of the employed filters was determined by taking the real (R) or the imaginary (I) part of the 1D Gabor filters. All spectral filters were assumed to have the same phase, and also all temporal filters were assumed to have the same phase, while spectral and temporal filters were allowed to have different phases. This structure allowed four SGBFB feature vectors with different combinations of spectral and temporal phases: Real-real (RR), real-imaginary (RI), imaginary-real (IR), and imaginary-imaginary (II) (cf. RR, RI, IR, and II in Fig. 5.3). To evaluate which of the phase combinations performs best in a robust ASR task, the four different SGBFB feature vectors were compared to GBFB and MFCC features on the CHiME (Computational Hearing in Multisource Environments) keyword recognition task. Barker et al. (2013) created the CHiME keyword recognition task to compare the robustness of ASR systems under controlled, realistic low-SNR conditions and to be able to compare the ASR performance to performance data from human listeners. Further, the role of the spectral and temporal modulation phase was assessed in recognition experiments combining several SGBFB feature vectors with different phase combinations.

5.2 METHODS

5.2.1 Spectro-temporal representation

The calculation of the LMSpec was based on an amplitude spectrogram with frames of 25 ms length and a temporal resolution of 100 frames/s. The linear frequency axis of the spectrogram was transformed to a Mel-scale using 31 equally spaced triangular filters with center frequencies

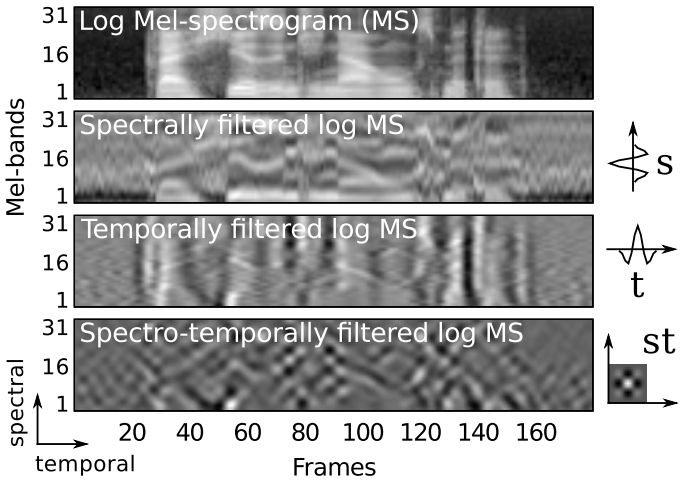

Figure 5.4: *Filtering a log Mel-spectrogram (LMSpec) by means of a 2D convolution: The LMSpec in the upper panel is 2D-convolved with a spectral 1D filter* s, *a temporal 1D filter* t *and the corresponding spectro-temporal 2D filter* st. *The result of the filtering process is depicted to the left of the corresponding filter. The amplitude of the 2D filters and (filtered) spectrograms is encoded in gray scale where white encodes high amplitude and black encodes low amplitude.*

in the range from 124 to 7284 Hz. The values of the amplitude Mel-spectrogram were subsequently converted to a decibel scale. All feature extraction schemes that are presented in the following extracted features from a LMSpec. An example of a LMSpec of a speech signal is depicted in the upper panel of Fig. 5.4.

5.2.2 Gabor filter bank features

2D GBFB features were extracted from a LMSpec using auditory-motivated spectro-temporal 2D Gabor filters, as described by Schädler et al. (2012a). There a LMSpec was 2D convolved (filtered) with a set of 2D Gabor filters to model the response of a range of neurons in the auditory cortex to the presented spectro-temporal patterns. The 2D filter shapes that were used to extract GBFB features are depicted in Fig. 5.1. These filters

were tuned to specific spectro-temporal modulation patterns that occur in speech signals and motivated by the fact that some neurons in the primary auditory cortex of mammals were found to be tuned to very similar spectro-temporal modulation patterns (Qiu et al., 2003). A 2D Gabor filter represents an idealized spectro-temporal receptive field and requires a pairing of spectral and temporal modulation frequencies. The pair of modulation frequencies determines a filter's shape and, hence, which spectro-temporal pattern would yield the strongest response in this particular filter. The main parameters of the employed 2D Gabor filters were the spectro-temporal center modulation frequencies and the spectral and temporal modulation bandwidths. Schädler et al. (2012a) structured the parameters of the 2D Gabor filters in a filter bank, which limited the number of free parameters and the correlation between the resulting feature dimensions. In this study, the same set of GBFB parameters was used, which was optimized for ASR and confirmed to extract robust ASR features (Moritz et al., 2013): The considered spectral modulation frequencies were $\omega_\mathrm{s} = 0.000$, 0.029, 0.060, 0.122, 0.250 cycles/channel. The considered temporal modulation frequencies were $\omega_\mathrm{t} = 0.0$, 6.2, 9.9, 15.7, 25.0 Hz. The number of half-waves under the envelope, which determines the bandwidth, in the spectral dimension was $\nu_\mathrm{s} = 3.5$. The number of half-waves under the envelope in the temporal dimension was $\nu_\mathrm{t} = 3.5$. The maximum extension of the filters in the spectral dimension was $b_\mathrm{s}^\mathrm{max} = 3 \cdot 31$, which is three times the number of Mel-bands. And the maximum extension of the filters in the temporal dimension was $b_\mathrm{t}^\mathrm{max} = 40$ frames (400 ms). The considered spectro-temporal center modulation frequencies were combinations of the spectral and temporal modulation frequencies and hence arranged on a grid (cf. Fig. 5.1). Spectral and temporal cross-sections through the maximum of the 2D frequency response of the GBFB filters with these parameters are shown in Fig. 5.5. To extract GBFB features from a LMSpec, it was convolved with each of the 41 2D Gabor filters, which resulted in 41 filtered LMSpecs. Subsequently, the filtered LMSpecs were spectrally sub-sampled at a rate of a quarter of the extent of the spectral width of the corresponding filter. This reduced redundancy from the filtered LMSpec, and was shown to be superior to using a Principle Component Analysis (Schädler et al., 2012a). The filtered and sub-sampled LMSpecs were concatenated and formed a 455-dimensional feature vector, which is referred to as GBFB features. The difference in dimensionality to the original GBFB features,

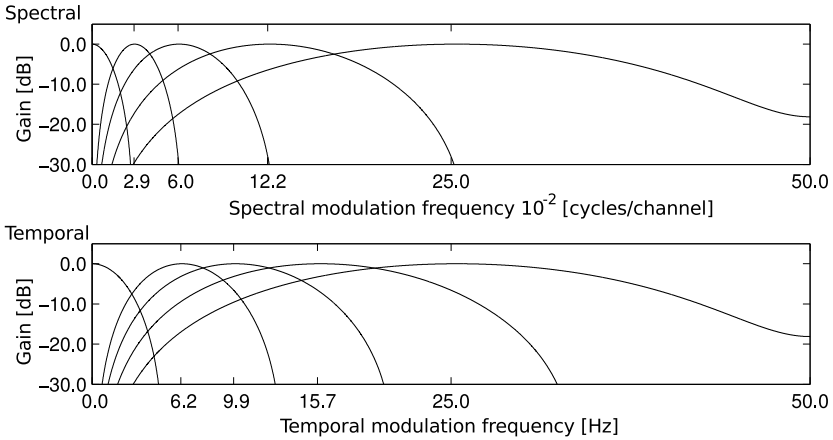

Figure 5.5: *Modified from Schädler et al. (2012a). Upper panel: Spectral cross-sections through the maximum of the 2D frequency response of GBFB filters; Lower panel: Temporal cross-sections through the maximum of the 2D frequency response of GBFB filters. The overlap of adjacent band-pass modulation filters is constant and governed by the distance between them and by their bandwidth.*

which are 311-dimensional, was due to the larger bandwidth (8 vs 4 kHz) that was considered in this study.

5.2.3 Separate Gabor filter bank features

Separate Gabor filter bank features (SGBFB) were extracted with two 1D Gabor filter banks, one spectral and one temporal, instead of with a filter bank of 2D Gabor filters.

5.2.3.1 1D Gabor filters

Equation (5.1) describes a 1D Gabor filter, where h_b is a Hann envelope function of width b, s_ω a sinusoid function with radian frequency ω, and g the product of both:

$$h_b(x) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi x}{b}\right) & -\frac{b}{2} < x < \frac{b}{2} \\ 0 & \text{else} \end{cases}, \tag{5.1a}$$

$$s_\omega(x) = \mathrm{e}^{\mathrm{i}\omega x}, \tag{5.1b}$$

$$g_{\omega,\nu}(x) = \underbrace{s_\omega(x)}_{\text{carrier}} \cdot \underbrace{h_{\frac{\nu}{2\omega}}(x)}_{\text{envelope}}. \tag{5.1c}$$

The width b is inversely proportional to the radian frequency ω and proportional to the number of half-waves under the envelope ν. Consequently, all 1D Gabor filters $g_{\omega,\nu}$ with the same value for ν are constant-Q complex-valued band-pass filters, where ω is the (radian) center frequency and determines the scale of the filter. The complex-valued filter function of a 1D Gabor filter with $\nu = 3.5$ half-waves under the envelope is depicted in Fig. 5.2, where E marks the absolute values (or envelope), R the real part, and I the imaginary part of the filter. Each of the different parts (E, R, and I) can be used to filter a signal. While E describes a low-pass filter, R and I are band-pass filters that only differ in phase and share the same frequency response. The width b of the Gabor filters is limited by $b^{\max}$. Filters with $\omega = 0$ would have an infinitely large support, which is why in this case the width of the envelope is set to $b^{\max}$, effectively resulting in a low pass filter (E). These filters (E, R, and I) can be applied in the spectral or in the temporal dimension to a LMSpec, resulting in a spectral or temporal modulation filtering, respectively. In

the following, a spectral filter bank and a temporal filter bank of 1D Gabor filters are presented.

5.2.3.2 1D Gabor filter banks

The center modulation frequencies (ω), the maximum filter width $b^{\max}$, and the number of half-waves under the envelope ν, which determines the filters' Q-factor were taken from the GBFB [cf. parameters from Schädler et al. (2012a)]. Hence the spectral modulation filter bank consisted of five 1D Gabor filters with $\nu = 3.5$, $b^{\max} = 93$ bands (three times the number of Mel-bands), and the following spectral modulation frequencies: ω = 0.000, 0.029, 0.060, 0.122, and 0.250 cycles/band. The temporal modulation filter bank consisted of five 1D Gabor filters with $\nu = 3.5$, $b^{\max} = 40$ frames, and the following spectral modulation frequencies: ω = 0.0, 6.2, 9.9, 15.7, and 25.0 Hz. As with GBFB filters, the envelope (E) function of width $b^{\max}$ was used as the filter function if the width of a filter function would exceed the maximum width $b^{\max}$, which here was the case for filters with $\omega = 0$. For all other filters ($\omega > 0$), only the real (R) or the imaginary (I) part of the filter was used as the filter function. As a result, in total, nine different spectral filters: 0.000 (E), 0.029 (R and I), 0.060 (R and I), 0.122 (R and I), and 0.250 (R and I) cycles/band, and nine different temporal filters: 0.0 (E), 6.2 (R and I), 9.9 (R and I), 15.7 (R and I), and 25.0 (R and I) Hz were considered. The real (R) part and the corresponding imaginary (I) part only differed in phase and hence shared the same frequency response. As a consequence, the frequency responses of the 1D spectral and 1D temporal Gabor filters were exactly the same as the cross-sections through the maximum of the 2D frequency responses of the 2D GBFB filters depicted in Fig. 5.5. Hence the two 1D Gabor filter banks covered the same range of spectral and temporal modulation frequencies as the 2D Gabor filters of the GBFB.

5.2.3.3 1D and 2D filtering of LMSpecs

The 1D filtering was performed by convolution with the corresponding filter functions. Temporal modulation filters were represented as row vectors and were convolved with each channel of the LMSpec independently. Likewise, spectral modulation filters were represented as column vectors and were convolved with each frame of the LMSpec independently. The temporal and spectral 1D filtering was performed by means of a 2D

convolution with row and column vectors, respectively. Therefore the LMSpec was convolved with a 1D row or column vector, as defined in Eq. (5.2), where k and n are the spectral and temporal indices of the LMSpec, respectively, and i and j the spectral and temporal offset of the filter from its center, respectively:

$$\begin{aligned}&\text{filtered-LMSpec}(k,n) :=\\&\sum_{i,j} \text{LMSpec}(k-i, n-j) \cdot \text{filterfunction}(i,j).\end{aligned} \tag{5.2}$$

filtered-LMSpec(k,n) was only calculated if LMSpec(k,n) existed, so that both the LMSpec and the filtered LMSpec, had the same size. In the following, a 2D convolution with a 1D filter, i.e., a filter the extent of which in the spectral dimension is one Mel-band or in the temporal dimension is one frame, is referred to as a 1D convolution or 1D filtering. Of course, a LMSpec can first be filtered spectrally, and the output can than be filtered temporally or vice versa. The order, i.e., if the spectral or temporal filtering is performed first, of this special form of spectro-temporal filtering does not affect the outcome. The outcome of a spectrally *and* temporally filtered LMSpec, is a spectro-temporally filtered LMSpec, and the corresponding spectro-temporal filter can be identified. In Eq. (5.3), a spectral filter s (column vector) and a temporal filter t (row vector) were applied in arbitrary order to a LMSpec:

$$\text{filtered-LMSpec} = [\text{LMSpec} * \text{s}] * \text{t}, \tag{5.3a}$$

$$= [\text{LMSpec} * \text{t}] * \text{s}, \tag{5.3b}$$

$$= \text{LMSpec} * \underbrace{[\text{s} * \text{t}]}_{\text{outer product: st}}, \tag{5.3c}$$

$$= \text{LMSpec} * \text{st}. \tag{5.3d}$$

In Eq. (5.3c), the 1D convolution with s and t was identified as the 2D convolution with the outer product of s and t. Hence the outer product of a spectral 1D and a temporal 1D filter is a *separable* filter because it can be described by independent spectral and temporal filter operations. The same is true for any 2D filter that can be described by a separate spectral and temporal 1D filter. Fig. 5.4 shows an example of a LMSpec of clean speech after filtering using temporal, spectral, and spectro-temporal filters. The corresponding filter functions are depicted to the right of the filtered LMSpecs.

5.2.3.4 Feature extraction

SGBFB features were extracted by first filtering the LMSpec spectrally, where either the R or the I phased filters were used, except for the DC filter ($\omega = 0$) for which always the E type was used. Due to the limited bandwidth in the output of spectral filtering processes with low center modulation frequencies, high correlations could be observed between some adjacent channels of the output. To the reduce these correlations, each spectrally filtered LMSpec was reduced in dimensionality by keeping only representative Mel-bands. This was achieved by critically sub-sampling the filtered LMSpec in spectral dimension at a rate of a quarter of the corresponding filters width b, where at least the center channel (Mel-band number 16), and at most all channels were kept. The same procedure for dimensionality reduction was used to extract GBFB features. The spectrally filtered and spectrally down-sampled LMSpecs were then filtered temporally, where either the R or the I phased filters were used, except for the DC filter ($\omega = 0$) for which always the E type was used. By the subsequent spectral and temporal filtering of the LMSpec, all considered spectral modulation frequencies were combined with all considered temporal modulation frequencies. The spectro-temporally filtered LMSpecs were concatenated and formed a 255-dimensional feature vector. These features are referred to as separate Gabor filter bank features or just SGBFB features.

With both the spectral and the temporal filter bank, the real (R) or the imaginary (I) part of the filters can be used. The filters that were actually employed are indicated by a suffix, where the first letter indicates the spectral and the second letter the temporal filter phase, e.g., SGBFB-RI. The effective spectro-temporal filter shapes for all possible combinations of all considered spectral and temporal E, R, and I filters are depicted in Fig. 5.6.

5.2.3.5 Spectro-temporal modulation phase

Because all four possible SGBFB feature vectors (SGBFB-RR, SGBFB-RI, SGBFB-IR, and SGBFB-II) covered the same range and combinations of spectral and temporal modulation frequencies and only differed in the phase of the modulation filters, it was investigated which phase combination offered the most robust representation in a speech-in-noise recognition experiment. Only two phase values were considered: The

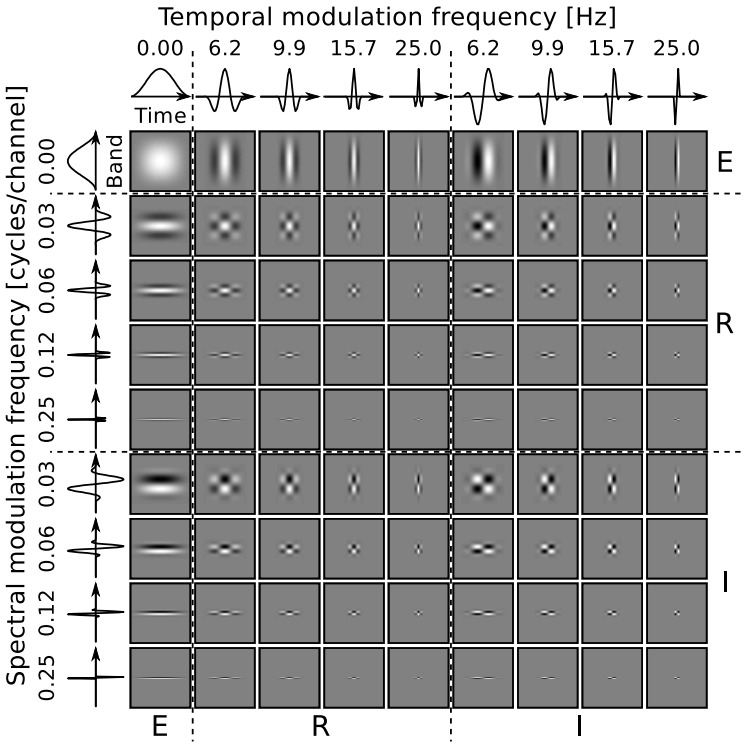

Figure 5.6: *All possible combinations of spectral and temporal 1D GBFB filters and their equivalent, separable spectro-temporal 2D filter functions. Each tile represents the outer product of the corresponding spectral and temporal filter functions with the horizontal axis within each tile being the temporal and the vertical axis being the spectral one. The 1D filters, depicted above and to the left of the 2D filters, are sorted by spectral and temporal center modulation frequencies, and are grouped according to the part of the complex 1D Gabor filter that is used: Envelope (E), real (R), imaginary (I). For a specific separate Gabor filter bank (SGBFB) feature vector, only a subset of these filters is used, which is indicated by a two-letter suffix. For example, for the SGBFB-RI feature set, the spectral E and R filters are combined with the temporal E and I filters. Note that each SGBFB feature vector covers the whole range of considered modulation frequencies, and that none of the* 81 *2D filter shapes is repeated.*

first one corresponded to the real (R) part of a Gabor filter (no phase shift), and the second one corresponded to the imaginary (I) part, where the carrier phase was shifted by $\pi/2$ rad relative to the real part. The real-real (RR) and imaginary-imaginary (II) spectro-temporal filters can be derived from the corresponding upward (U) and downward (D) filters by addition (A) and subtraction (S) as depicted in Fig. 5.3, while the real-imaginary (RI) and imaginary-real (IR) phase combinations cannot be represented by linear combination of any two filters of the 2D GBFB. To take multiple phase combinations into account, different *single* SGBFB feature vectors were concatenated and the robustness of the combined—or *dual*—SGBFB feature vectors was determined in a speech-in-noise recognition experiment. The concatenation of two 255-dimensional, single feature vectors resulted in a 510-dimensional dual feature vector and is referred to as SGBFB-X-Y, where X determined the phases of the first and Y the phases of the second vector, e.g., SGBFB-RR-II. A dual SGBFB feature vector represented all spectro-temporal modulation frequencies twice, in contrast to the 455-dimensional GBFB feature vector, where only the modulation frequencies of the truly spectro-temporal filters were represented twice [cf. upward (U) and downward (D) filters in Fig. 5.1]. This explains the difference in dimensionality between dual SGBFB feature vectors and the GBFB feature vector. The concatenation of feature vectors with all possible phase combinations combined all considered spectral and temporal 1D filters and hence extracted 1020-dimensional feature vectors effectively using all 81 2D patterns depicted in Fig. 5.6. These feature vectors are referred to as *complete* SGBFB features or SGBFB-RR-RI-IR-II.

5.2.4 Feature normalization

Blind feature statistics adaptation, such as mean and variance normalization (MVN) (Viikki and Laurila, 1998) or histogram equalization (HEQ) (De La Torre et al., 2005) can improve the robustness of an ASR system. All features were normalized using histogram equalization (HEQ). As each feature dimension was processed independently, the process is only described for one feature dimension, which is considered to be a time series. While mean and variance normalization normalizes the first two statistical moments of the distribution of the values of the time series, HEQ can normalize even higher statistical moments, such as skewness and kurtosis. For this, the values of the time series were projected by a

function that mapped the source distribution to a desired target distribution. The mapping function was estimated by calculating 100 percentiles (e.g., 0.5%, 1.5%, ..., 99.5%) of the source distribution and mapping these to the same percentiles of the desired target distribution, where values between the percentiles were interpolated linearly. Care needed to be taken when estimating the percentiles of the source distribution, as the 0% and 100% percentiles could not be reached with finite time series. The maximum expected percentile $\mathrm{p}_N^{\max}$ and minimum expected percentile $\mathrm{p}_N^{\min}$ when drawing N samples from a distribution were estimated using Eq. (5.4):

$$\mathrm{p}_N^{\max} = 100 * \frac{N}{N+1}, \tag{5.4a}$$

$$\mathrm{p}_N^{\min} = 100 * \frac{1}{N+1}. \tag{5.4b}$$

Therefore, 100 equally spaced percentiles between $\mathrm{p}_N^{\min}$ and $\mathrm{p}_N^{\max}$ were mapped to the corresponding percentiles of the standard normal distribution, where N was the number of feature vectors. The resulting time series had—within the limits due to mapping only 100 percentiles—the same moments as the standard normal distribution. All features were processed with HEQ on a per-utterance basis, where the average utterances length of the employed corpus was 1.8 ± 0.25 s.

5.2.5 Recognition experiment

The task that was employed to evaluate the robustness of ASR systems is the recognition of English commands being spoken in noisy living room environments that were recorded using an binaural manikin. Therefore the training, development, and test data sets from the first track of the second CHiME challenge (Vincent et al., 2013a) were used. The sentences of this corpus were recorded from 34 different (male and female) speakers. They have a fixed syntax of the form "command color preposition letter number adverb" (e.g., "put red at G9 now"), where the words were drawn from a closed vocabulary. The utterances of the development and test data set were filtered with the binaural combined head and room impulse responses of two rooms (a lounge and a kitchen) corresponding to a frontal position at a distance of 2 m. Subsequently, they were mixed with noise samples recorded using the binaural manikin in the same environments at SNRs from -6 to 9 dB. In this study, the binaural signals were mixed

down to one channel prior to the feature extraction by adding the left and the right channel. The whole sentences had to be recognized but only the percent correct value of the letter (in the example: G) and the digit (in the example: 9) was evaluated as in the first track of the second CHiME challenge.

Three different training data sets were available and used to evaluate the performance of ASR systems depending on the training condition: Clean, reverberated, and isolated (which is noisy and reverberated). While the clean data set contained unprocessed speech samples, the utterances of the reverberated and isolated data sets were filtered with the binaural impulse responses. The utterances of the isolated (or noisy) data set were additionally mixed with noise samples that were recorded with the binaural manikin in the corresponding room at SNRs from −6 to 9 dB. Even though some of the considered front-ends might have performed better with additional training data, the unmodified training data sets from the CHiME challenge were used for the sake of comparability. For evaluation, each ASR system was trained with the three different training data sets. While all pilot experiments had been conducted with the development data set, the results were obtained on the test data set.

The training and testing scripts provided in the CHiME challenge are based on HTK (Young et al., 2006). The differences between the provided scripts and the scripts that were actually used for conducting the experiments are highlighted in Sec. 5.2.7. For each training data set, the recognition performance in percent-of-digits-and-letters correct was measured at SNRs from −6 to 9 dB in 3 dB steps. The uncertainty of the performance measure due to the limited amount of test sentences (600) was estimated in advance, because it consisted of 1200 independent binary decisions; 600 for digits and 600 for letters. At 50% correct it happened to be about 1.45 percentage points, at 70% correct about 1.32 percentage points, and at 90% correct it was estimated to be about 0.85 percentage points. The recognition results, which depend on the SNR, were compared between different systems by calculating the relative change in SNR that would be required to get the same performance with two different systems, as described in Sec. 5.2.6. Additionally, human recognition performance data from the first CHiME challenge was available and used to present selected results in terms of the remaining *man-machine gap*, as described in Sec. 5.2.8.

5.2.6 Robustness measure

To report the relative improvement of a system over a reference system in a single value with physical meaning, the equal-performance increase in dB SNR (EPSI) is reported. This type of reporting is related to the speech reception threshold, which is widely used to measure the performance of human listeners to recognize speech in noise. The speech reception threshold is the SNR that is required to understand a specific portion, e.g., 50%, of the presented speech material. To use all available data points, the comparison was carried out at different performance levels. Hence the difference in SNR between the performances of two recognition system was integrated over the performance range where two systems could be compared. Let $P(r)$ be the performance graph of an ASR system, with r being the SNR in dB and P being the recognition performance at that SNR. Applying Eq. (5.5) guarantees the monotonicity of the performance graphs $P^{\mathrm{mon}}(r)$:

$$P^{\mathrm{mon}}(\mathrm{SNR}) = \min_{r \geq \mathrm{SNR}} P(r). \tag{5.5}$$

The performance levels at which the systems were compared were interpolated in 0.5 dB steps in the region that data for both systems was available, as illustrated in Fig. 5.7. The average over the differences in SNR is invariant under any monotonic transformation of the performance axis. It is intuitively interpreted as the increase (or decrease) in SNR that is needed to get the same performance with the compared system as with the reference system. When comparing two ASR systems A and B, a symmetric EPSI was achieved by averaging the differences with A as the reference for B and with B as the reference for A. Ideally, the recognition performance of human normal-hearing listeners would have been used as a reference for all experiments. Although human performance data existed for the employed task, the human speech recognition (HSR) performance at the lowest SNR (-6 dB) was about 90% word recognition rate; so good that only few ASR systems could have been compared to it. Hence a reference ASR system was used instead, and only the best performing systems were compared to HSR performance.

5.2.7 Reference systems

Standard MFCCs and GBFB features with HEQ served as standard reference features. MFCCs were extracted from a LMSpec by spectrally

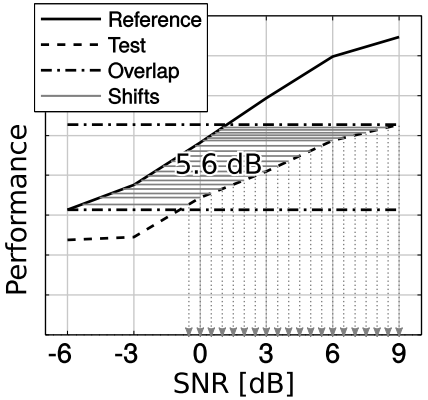

Figure 5.7: *Illustration of comparing the robustness of two ASR systems in terms of changes in signal-to-noise ratio (SNR). The average relative increase/shift of the SNR for a test ASR system that is required to achieve equal performance with a reference system can be calculated independently from the scaling of the performance axis. Therefore the integration points are selected on the SNR axis in* 0.5 *dB steps in the range where the performance graphs overlap on the performance axis.*

processing it with a discrete cosine transform, where only the first 18 coefficients, which account for spectral modulation frequencies from 0 to 0.29 cycles/channel, were used. The 18 MFCCs were concatenated with their first and second discrete temporal derivatives, which were calculated by applying a temporal slope filter of five frames length once or twice, respectively. The resulting MFCC feature vector, which included both derivatives, was 54-dimensional. The extraction of the 455-dimensional GBFB feature vectors is described in detail in Sec. 5.2.2. All features that were evaluated in this study were normalized using HEQ as described in Sec. 5.2.4.

On the back-end side, GMM and HMM were used to model speech. The training and testing scripts provided in the first CHiME challenge (Barker et al., 2013) are based on the Hidden Markov Toolkit (HTK) (Young et al., 2006). Deviating from the default configuration, the reference system used tri-phone models instead of whole-word models. The required changes to the training procedure were based on the HTK Wall Street

Journal Training Recipe from (Vertanen, 2006). Three-state left-to-right tri-phone speaker-depended acoustic models, a three-state background model with skip and back transitions, and a one-state short pause model tied to the center state of the background model were employed. The CMU Pronouncing Dictionary (Carnegie Mellon University, 2007), version 7a, was employed to generate initial monophone labels, where an optional short pause was allowed between two words. After the initial training of speaker-independent monophone models, tri-phone models of all possible monophone combinations were generated and initialized with the model of the center monophone. The parameters of the tri-phones were re-estimated in four iterations and subsequently tied with tri-phones that share the same center monophone using HTK's tree-based state tying method. The decision tree phonetic questions that are needed for the tree-based state tying were taken from Vertanen (2006). The threshold that governs the number of tied states was chosen so that the number of tied states was 700 ± 2. The number of Gaussian mixture components per state was increased stepwise to 2, 3, 5, and 7 in the course of the training procedure, with four iterations of parameter re-estimation in-between. The models were then adapted to the speaker using HTK's maximum *a posteriori* (MAP) method to update the mean values and the mixture weights, instead of using HTK's parameter re-estimation. The recognition of utterances was performed with the corresponding speaker-dependent model, where a language model enforced the syntax of recognized sentences (command color preposition letter number adverb).

5.2.8 Man-machine gap

To put the results of this study into the perspective of building an ASR system that is as robust as a normal-hearing human listener, selected results were compared to literature data of HSR performance, which is available from the first CHiME challenge (Barker et al., 2013). The difference between the first CHiME challenge and the first track of the second CHiME challenge is that in the latter head movements of the speaker are simulated, which we consider to have a negligible effect on the HSR data for our purposes. The equal-performance increase in dB SNR (EPSI) of the ASR over the HSR results was used to quantify the remaining *man-machine gap*. In addition, the results for a GBFB-based system from the literature, which was presented by Moritz et al. (2013) during the second CHiME keyword recognition challenge and placed

second, were also compared. This system, referred to as GBFB-CC, exploited binaural information using source separation based on non-negative matrix factorization, and featured a more sophisticated speaker adaptation, which includes in addition a maximum likelihood linear regression (MLLR) parameter adaptation step.

5.2.9 Reference implementations

MATLAB reference implementations of several methods, including the calculation of the LMSpec, MFCC features, GBFB features, SGBFB features, the HEQ, and the EPSI, are available online[1].

5.3 RESULTS

All evaluated features sets were normalized using HEQ, as described in Sec. 5.2.4, and evaluated on the CHiME keyword-in-noise recognition task, as described in Sec. 5.2.5, using the ASR system described in Sec. 5.2.7, where the reference features were replaced with the features in question. The relative improvements are reported in EPSI, which is defined in Sec. 5.2.6. The uncertainty of all results was propagated from the estimated uncertainty due to the limited number of test sentences, as explained in Sec. 5.2.5.

5.3.1 Performance of reference system and data representation

The absolute recognition scores of the reference systems along with the approximate HSR performance depending on the SNR in decibels are depicted in Fig. 5.8, and reported in numerical form in Table 5.1. As expected, the human performance was found to be superior to the performance of the ASR systems. Independent of the used features, the ASR systems that were trained on the noisy data set performed better at lower SNRs (less than 3 dB), while for high SNRs, particularly at 9 dB SNR, the systems trained on only reverberated data performed better. The ASR systems that were trained on the clean data set performed much worse, which is why these results were not considered to be a good indicator for robustness. Because we were interested in noise robustness,

[1] URL: http://medi.uni-oldenburg.de/SGBFB

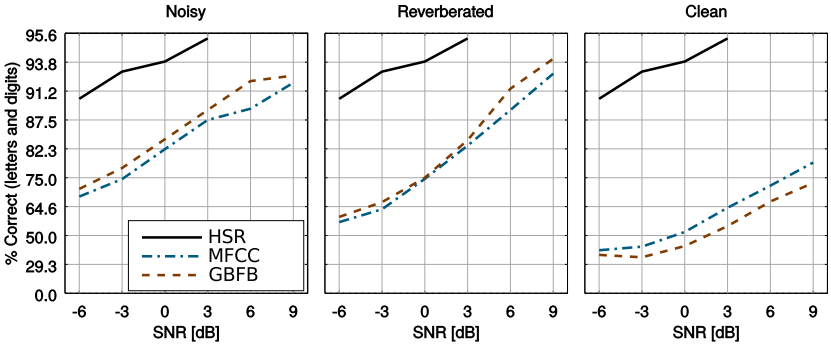

Figure 5.8: *Recognition performance on the test data set of the MFCC and GBFB-based reference ASR systems depending on the SNR and the training data set, along with the approximate human speech recognition (HSR) performance. The word recognition rate in percent correct is plotted over the test SNR for systems trained with* clean, reverberated, *and* noisy *speech data. The y axis is a logarithmically scaled word error rate axis, which is labeled with the corresponding word correct rates in percent.*

Table 5.1: *Recognition performance of the MFCC and GBFB-based reference ASR systems on the second CHiME keyword-in-noise recognition task in percent correct along with the human speech recognition (HSR) performance, which was measured during the first CHiME challenge. The systems were trained with clean, reverberated, or noisy data, and evaluated on the noisy test data set.*

Features	Train condition	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB
HSR	–	90.3	93.0	93.8	95.3	96.8	98.8
MFCC	Clean	40.3	42.8	52.1	64.2	72.5	79.2
MFCC	Reverberated	57.4	63.5	74.7	83.0	88.9	92.8
MFCC	Noisy	68.7	74.6	82.2	87.5	89.1	92.0
GBFB	Clean	36.9	35.1	43.2	55.3	66.8	73.4
GBFB	Reverberated	60.0	66.5	75.0	84.1	91.4	94.0
GBFB	Noisy	71.4	77.8	84.2	88.9	92.2	92.7

Table 5.2: *Equal-performance increase in dB SNR (EPSI) of the reference ASR systems over HSR data for different training conditions, where a value of X means the SNR needs to be increased by X on average for the corre- sponding system to perform as well as a human listener. Using GBFB fea- tures reduced the distance to human performance compared to when using MFCCs from* 13.2 *dB to* 10.6 *dB SNR.*

Features	Noisy	Reverberated	Clean
MFCC	$+13.2 \pm 0.95$	$+12.6 \pm 1.00$	–
GBFB	$+10.6 \pm 1.12$	$+10.3 \pm 1.06$	–

not the ability of generalizing from quiet to noisy conditions, the results for ASR systems trained with noisy data were taken as the indicator of robustness. To compare the ASR system with different features regarding their robustness on the CHiME task, the EPSI measure presented in Sec. 5.2.6 was used to report the difference in performance in a single, physically interpretable value; the equal-performance increase of the SNR in decibels. The EPSIs of the reference ASR systems over the HSR performance in dB are reported in Table 5.2. The MFCC-based reference system required the SNR to be $+13.2 \pm 0.95$ dB higher to perform as well as an average native human listener, while the GBFB-based reference system required the SNR only to be $+10.6 \pm 1.12$ dB higher. Hence the GBFB-based system was found to be more robust than the MFCC based system on this task. For ASR systems that do not reach HSR performance, such as the systems trained with clean data, the EPSI over HSR performance cannot be calculated. This is the reason why in the following the GBFB-based reference system was used as the baseline for the comparison.

5.3.2 Single SGBFB features

Table 5.3 reports the EPSIs of the differently phased 255-dimensonal SGBFB features over the GBFB reference system. We considered the results for the noisy training condition to carry the most information about the features' ability to facilitate the back-end of the recognition of speech in noise. The relative increase in SNR to achieve equal performance for the clean and reverberated training condition are reported for completeness. The SGBFB-IR system, which uses the imaginary part of the 1D Gabor

Table 5.3: *Average equal-performance increase in dB SNR over the GBFB reference system to achieve the same performance with single SGBFB features when training on clean, reverberated, or noisy data. A positive value indicates that the system under consideration performs worse than the GBFB reference system.*

Features	Noisy	Reverberated	Clean
SGBFB-RR	$+2.2 \pm 0.45$	$+0.7 \pm 0.30$	-0.7 ± 0.36
SGBFB-RI	$+2.7 \pm 0.45$	$+0.8 \pm 0.27$	-1.1 ± 0.31
SGBFB-IR	$+1.1 \pm 0.44$	$+1.5 \pm 0.26$	-0.0 ± 0.30
SGBFB-II	$+2.5 \pm 0.42$	$+1.7 \pm 0.28$	-0.9 ± 0.35

Table 5.4: *Average equal-performance increase in dB SNR over the GBFB reference system for ASR systems with MFCC or dual SGBFB features when being trained on clean, reverberated or noisy data.*

System	Noisy	Reverb	Clean
SGBFB-RR-RI	-0.3 ± 0.46	$+0.4 \pm 0.28$	$+0.7 \pm 0.30$
SGBFB-RR-IR	$+1.2 \pm 0.42$	$+0.7 \pm 0.28$	-0.1 ± 0.34
SGBFB-RR-II	-0.7 ± 0.47	-0.0 ± 0.29	-0.5 ± 0.31
SGBFB-RI-IR	-0.9 ± 0.45	$+0.1 \pm 0.29$	-1.0 ± 0.35
SGBFB-RI-II	$+1.8 \pm 0.43$	$+0.7 \pm 0.28$	-1.7 ± 0.31
SGBFB-IR-II	-0.4 ± 0.43	$+0.6 \pm 0.28$	-1.0 ± 0.36

filter for spectral filtering and the real part for temporal filtering, is the one that came closest to the GBFB reference with a EPSI of $+1.1 \pm 0.44$ dB. This means that the ASR system with SGBFB-IR features required the SNR to be 1.1 ± 0.44 dB higher than with GBFB features to get the same performance. With the other SGBFB features, the EPSI increased to more than 2 dB. The ASR system with GBFB features outperformed all ASR systems using only single SGBFB feature vectors or MFCCs.

5.3.3 Dual SGBFB features

The required increase in dB SNR for all ASR systems using dual SGBFB feature vectors, which are combinations of two differently phased single SGBFB feature vectors, to achieve equal performance with the GBFB reference system are reported Table 5.4. The best dual SGBFB feature

set was the one that concatenates SGBFB-RI and SGBFB-IR feature vectors to 510-dimensional SGBFB-RI-IR feature vectors. It yielded an improvement over the GBFB reference of -0.9 ± 0.45 dB, i.e., a decrease in SNR to achieve the same performance. In terms of word error rates, this translates to an average relative improvement of 8.3% over the GBFB reference system, and 20.6% over the MFCC reference system, where an improvement of 50% would correspond to halving the word error rate. The dual SGBFB feature vectors with the *same temporal phase* and different spectral phases (RR-IR, RI-II) performed worse than the GBFB reference. Those with the same spectral phase and *different temporal phases* (IR-II, RR-RI) performed as well as GBFB features within the uncertainty imposed by the setup. Those with *different spectral and temporal phases* (RI-IR, RR-II) improved the robustness of the GBFB-based reference system.

Using the MATLAB reference implementation, the 2D GBFB spectro-temporal filtering achieved a real-time factor of 0.4887 (median of 100 runs), while the 1D SGBFB-RI-IR spectro-temporal filtering achieved a real time factor of 0.0078 (median of 100 runs) on the same PC system[2], i.e., the separate processing was found to be about 60 times faster. Hence by using dual SGBFB features instead of GBFB features, the computational time required for the spectro-temporal filtering was reduced by more than an order of magnitude, while at the same time the robustness was increased.

5.3.4 Complete SGBFB features

When concatenating all differently phased SGBFB features to 1020-dimensional SGBFB-RR-RI-IR-II feature vectors, the EPSI over the GBFB reference was -1.2 ± 0.42 dB when training on noisy data. In terms of word error rates, this translates to an average relative improvement of 12.8% over the GBFB reference system, and 24.8% over the MFCC reference system, where 50% would mean halving the word error rate. The most robust front-end evaluated in this study was the complete SGBFB feature set.

[2] CPU: AMD A10 PRO-7350B @ 2.1 GHz; RAM: DDR3L-1600 CL9; Matlab R2010b (single thread) on Linux 3.13

5.3.5 Quantity of training data

A reasonable question when using ASR systems with high-dimensional features is whether sufficient training data are available because the number of GMM parameters increases proportionally with the number of feature dimensions. On the one hand, using scarce training data could favor systems that require less parameters to be determined during the training phase and prevent systems with more parameters from showing their full potential. On the other hand, using large amounts of training data could conceal the possibility that systems using high-dimensional features might require these amounts of data, while systems with low-dimensional features would not perform worse when using less training data. To test if one or the other was the case, systems with the low-dimensional MFCC features and the high-dimensional SGBFB-RI-IR features were trained with a reduced training data set, which contained only half of the training sentences that were available per speaker, i.e., 250 instead of 500. With this reduced training data set, the system that uses the 54-dimensional MFCC features performed 2.2 ± 0.44 dB (EPSI) worse and the system that uses the 510-dimensional SGBFB-RI-IR features performed 2.0 ± 0.46 dB worse compared to when using the full training data set. This result shows that the systems with high- and low-dimensional features were equally affected when the amount of training data was halved, and hence that no system was favored due to the amount of training data that were used in the recognition experiments. Compared to the system with MFCC features that was trained on the full training data set, the system with SGBFB-RI-IR features that was trained with the reduced training data set performed about (± 0.5 dB) the same. Hence we are confident that the training data set from the CHiME challenge provided a fair comparison of the differently-dimensional feature sets.

5.3.6 Remaining man-machine gap

Figure 5.9 depicts the absolute word recognition rates of the reference systems, the best SGBFB system, the GBFB-CC system, and from HSR experiments. Table 5.5 reports the EPSIs over human speech recognition performance that quantify the remaining man-machine gap. While the MFCC-based reference ASR system required the SNR to be about 13 dB higher to perform as well as a human listener, the GBFB-based reference

system still had an EPSI of about 11 dB, and the best SGBFB-based system one of about 9 dB. Hence the gap in speech recognition robustness between man and machine remains but was reduced by 2 dB by using SGBFB features instead of GBFB features.

5.4 DISCUSSION

5.4.1 Modulation phases

The main results reported in Tables 5.3 and 5.4 indicate that an ASR system with a combination of SGBFB features may exhibit a greater robustness than the GBFB reference system if the phase of the spectral and temporal modulation filters is chosen in an appropriate way. The ASR systems with single SGBFB features vectors (RR, RI, IR, and II), which consider only one spectral and one temporal phase constellation, were found to be less robust than the GBFB reference system, where

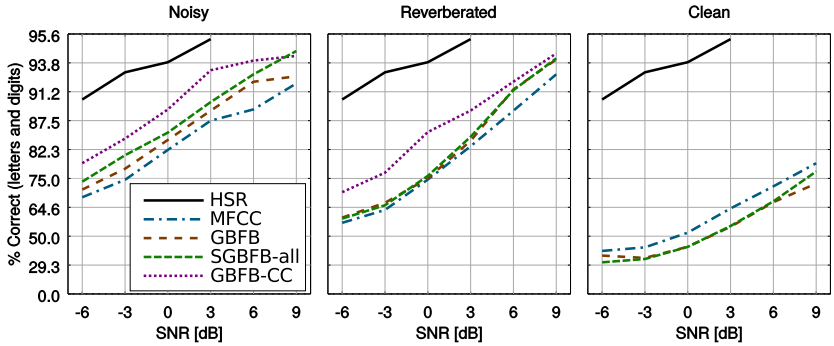

Figure 5.9: *Recognition performance on the test data set of different ASR systems and human speech recognition (HSR) experiments depending on the SNR and training data set. Besides the performance of ASR systems using MFCC features, GBFB features, or the complete SGBFB feature set (SGBFB-all), the performance of a GBFB-based system with binaural processing (GBFB-CC) from Moritz et al. (2013) from the second CHiME challenge is depicted. The word recognition rate in percent is plotted over the test SNR. The y axis is a logarithmically scaled word error axis, which is labeled with the corresponding word correct rates in percent correct.*

Table 5.5: *Equal-performance increase in SNR over HSR performance in dB for different training conditions, where a value of X means the SNR needs to be increased by X on average for the corresponding system to perform as well as a human listener. Using GBFB features reduces the distance to human performance compared to when using MFCCs from* 13.2 *to* 10.6 *dB SNR. The use of dual SGBFB features can reduce the distance to* 9.5 *dB, and the use of all SGBFB feature vectors combined can reduce the distance further to* 8.6 *dB. The GBFB-based system from Moritz et al. (2013) which, like humans and opposed to the other systems, exploits binaural information (GBFB-CC), even gets as near as* 6.2 *dB to human performance.*

System	Isolated	Reverb	Clean
MFCC	+13.2±0.95	+12.6±1.00	–
GBFB	+10.6±1.12	+10.3±1.06	–
GBFB-CC	+6.2±1.19	+9.4±1.04	–
SGBFB-RI-IR	+9.5±1.19	+10.1±1.04	–
SGBFB-RR-RI-IR-II	+8.6±1.01	+10.2±1.09	–

the systems with real-phase temporal filters (I**R** and R**R**) performed better than those with imaginary-phase temporal filters (I**I** and R**I**). To build a system with SGBFB features that was at least as robust as the reference system with GBFB features, a dual SGBFB feature vector with both temporal phase constellations was required (R**R**-R**I**, R**R**-I**I**, R**I**-I**R**, and I**R**-I**I**). If the temporal phase was the same (R**R**-I**R** or R**I**-I**I**), the corresponding system performed worse than the GBFB reference system. To improve the robustness of the GBFB reference system, both temporal and both spectral phase constellations were required (**RR-II** and **RI-IR**). Finally, the ASR system using complete SGBFB features, which include all possible phase combinations (RR-RI-IR-II), was found to be the most robust one. These findings suggest that the temporal phase is more important than the spectral phase and that diverse phase information of modulation filters is beneficial to the robustness of ASR systems.

The phase of the modulation filters was found to be an important factor. However, it does not affect the frequency response of the filters, which indicates that modulation filters in the context of robust ASR are insufficiently described by only specifying their frequency response. A reason that considering temporal and spectral modulation filters with

orthogonal, shifted carrier functions (i.e., the real and the imaginary part) benefits the robustness of ASR systems could be that their output is not systematically correlated, which is a property that GMMs with diagonal covariance matrices are well-disposed to. For example, for the temporal domain, the shape of the imaginary filter (I) is very similar to the shape of the slope (or delta) filter, which is traditionally used to calculate the first discrete temporal derivative with MFCCs, and the shape of the real filter (R) is very similar to the shape of the double-delta filter, which is traditionally used to calculate the second temporal derivative. Both describe different properties and seem to encode complementary information, which is why the combination of differently-phased feature vectors could improve the robustness. But while with the delta filters only one temporal center frequency was extracted, with the SGBFB filters considered here, five different modulation frequencies between 0 and 25 Hz were extracted.

While the whole spectral context was always available to the back-end in the same feature vector, the temporal context was distributed over several feature vectors. A reason why the temporal phase was found to be more important than the spectral phase in this regard could be that the HMM back-end is inherently probabilistic about *timing* and could have benefited from the presence of additional hard-coded temporal information in the feature vectors. The effect of changing the temporal phase is that the carrier is shifted in time, while the window function (the envelope) remains invariant. The output of the temporal filters with shifted carriers could have conveyed information that otherwise was not accessible to the back-end.

5.4.2 1D vs 2D Gabor filter complexity

Separating the spectro-temporal 2D GBFB into two SGBFB was not only found to improve the robustness of an ASR system in difficult acoustic conditions but also to achieve this with less complex filters. While with the 2D GBFB filters the spectral filtering and the temporal filtering are dependent and happen simultaneously, with the 1D SGBFB filters, the spectral and the temporal filtering are independent and can be carried out in arbitrary order. This reduces the complexity of the features and also of the feature calculation because no spectro-temporal interactions need to be considered. The corresponding reduction in computational time, that was required for the spectro-temporal processing, was found

to be more than an order of magnitude. It is yet to be investigated if the 1D Gabor filters and the chosen parameter values for the filter width and center modulation frequencies are the optimal choice for robust ASR. But, at least in the studied context, it seems that truly spectro-temporal filters did not give an advantage over separate spectro-temporal filters. This suggests that future research on robust speech features might reasonably assume spectro-temporal interactions (such as, e.g., temporal changes of spectral information as in glides or formant transitions) to play a minor role in comparison to having both temporal and spectral information available simultaneously.

5.4.3 Remaining man-machine gap

A part of the remaining gap between the complete SGBFB feature based system (SGBFB-RR-RI-IR-II) and the HSR performance in Table 5.5 could be due to the very basic binaural processing (down-mixing) that was employed in this study, which did not exploit binaural cues for noise reduction as opposed to the human auditory system and the GBFB based system from the chime challenge (GBFB-CC). A SGBFB based system that exploits binaural information could provide further improvements in robustness. Another, maybe even related, reason could be the negligence of any phase—not modulation phase—information of the spectral channels. The temporal fine structure, which encodes binaural information as well as information about voicing or the harmonic structure of a signal, is not considered at all when using a LMSpec as a basis for feature extraction. This information could help to group signal parts and better separate them from the rest. The current research on this topic in the field of computational acoustic scene analysis (CASA) might some day converge with the investigation on robust speech recognition. For now, the SGBFB feature extraction algorithm permits the investigation of spectral and temporal modulation processing independently and to assess the interdependence of both types of processing in the context of speech recognition.

Even though the omission of certain modulation frequencies or spectro-temporal modulation pairs might be a good tool to systematically evaluate the relative importance of these features, this endeavor was beyond the scope of this paper and might be considered in future work.

5.5 CONCLUSIONS

The most important findings of this work can be summarized as follows:

— A combination of separate spectral and temporal 1D Gabor modulation filter banks (SGBFB) was successfully employed instead of the spectro-temporal 2D GBFB to extract robust ASR features. SGBFB features improved the robustness over GBFB features by up to 1.2 dB SNR, which corresponds to an average relative improvement of the word error rate of 12.8% over a GBFB based reference system, and 24.8% over a MFCC based reference system.

— While a close interaction between temporal and spectral processing was found to be comparatively irrelevant for robust ASR, the phase of the spectral and especially the temporal modulation filters was found to be an important factor, which can be used to provide complementary and additional temporal information to the back-end.

— Compared to human listeners, the SNR needed to be 13 dB higher for a MFCC-based system, 11 dB higher for a GBFB-based, and 9 dB higher for a SGBFB-based system, to achieve the same recognition performance.

ACKNOWLEDGMENTS

This work was supported by the DFG (SFB/TRR 31 “The active auditory system”). We would like to thank Bernd Meyer for his support and contribution to this work.

6 Matrix sentence intelligibility prediction using an automatic speech recognition system

ABSTRACT

Objective: The feasibility of predicting the outcome of the German matrix sentence test for different types of stationary background noise using an automatic speech recognition (ASR) system was studied. *Design:* Speech reception thresholds (SRT) of 50% intelligibility were predicted in seven noise conditions. The ASR system used Mel-frequency cepstral coefficients as a front-end and employed whole-word Hidden Markov models on the back-end side. The ASR system was trained and tested with noisy matrix sentences on a broad range of signal-to-noise ratios. *Study sample:* The ASR-based predictions were compared to data from the literature (Hochmuth et al., 2015) obtained with 10 native German listeners with normal hearing and predictions of the speech intelligibility index (SII). *Results:* The ASR-based predictions showed a high and significant correlation ($R^2 = 0.95, p < 0.001$) with the empirical data across different noise conditions, outperforming the SII-based predictions which showed no correlation with the empirical data ($R^2 = 0.00, p = 0.987$). *Conclusions:* The SRTs for the German matrix test for listeners with normal hearing in different stationary noise conditions could well be predicted based on the acoustical properties of the speech and noise signals. Minimum assumptions were made about human speech processing already incorporated in a reference-free ordinary ASR system.

This chapter is a reformatted reprint of "Matrix sentence intelligibility prediction using an automatic speech recognition system", M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, International Journal of Audiology Volume 54:2 pp. 100–107, which was published the 18th of September 2015 by Taylor & Francis Ltd (www.tandfonline.com). Reprinted by permission of the publisher. The original article can be found at http://dx.doi.org/10.3109/14992027.2015.1061708.

6.1 INTRODUCTION

Accurate speech intelligibility predictions are of great importance for various practical applications such as the objective evaluation of different signal processing strategies in hearing assistive devices, or room acoustic design. If appropriately adjusted, they can also be used as a tool to assess the effect of different aspects of hearing impairment on speech intelligibility in an objective way. Several models were proposed to predict speech intelligibility in noise including static measures (Articulation index / speech intelligibility index(SII); Kryter (1962); ANSI (1970, 1997), temporal measures (speech transmission index, Steeneken and Houtgast, 1980), measures based on automatic speech recognition (Cooke, 2006), or measures taking psychoacoustical or physiological aspects of the auditory periphery into account (Holube and Kollmeier, 1996; Stadler et al., 2007; Jürgens and Brand, 2009; Jørgensen et al., 2013). These models usually make strong assumptions about the relation between the (time-dependent) signal-to-noise ratio (SNR) in frequency bands and the resulting speech intelligibility, or about the processing steps in the auditory system and the auditory pattern recognition process. Any deviation between predictions and real human data can usually be attributed to the failure of one or several of these assumptions. Since the accuracy of predicting speech intelligibility even for listeners with normal hearing in different listening conditions (such as different types of background noise) is very limited (e.g., Hochmuth et al., 2015), the validity and relevance of these assumptions is unclear.

The goal of the present study was to overcome these limitations by accurately predicting the performance of listeners with normal hearing in different stationary noise conditions with a minimum set of assumptions. This should help to assess the relevance of some of the current assumptions used in models of human speech intelligibility and to provide a valid baseline for speech intelligibility modeling with possible applications in audiology and acoustic communication research. For this, we employed an ordinary automatic speech recognition (ASR) system to predict the outcome of the German matrix test using data available from the literature.

The closed-set matrix sentence test, first proposed by Hagerman (1982), has primarily been developed for hearing assessment and is widely used in clinical practice and research in different languages (see review by

Kollmeier et al., 2015). It uses sentences with a fixed syntax (name-verb-number-adjective-object), but virtually no semantic meaning, like "Peter sees eight wet chairs". However, this test structure is well suited for the purpose of the current paper due to its limited vocabulary and its fixed syntactic structure which facilitates the training of an appropriate ASR system (see below). The advantage of this approach is the direct modeling of the speech perception process without making as many assumptions and simplification as with, e.g. the SII, which is one of the most commonly used objective measures for human speech recognition.

The SII is based on early findings (Fletcher and Galt, 1950) that human speech intelligibility depends on the proportion of spectral information that is audible to the listener. It is computed by dividing the spectrum of speech and noise separately into frequency bands and estimating the weighted average of the band-specific signal-to-noise ratios. The weighting reflects band-importance functions that were determined for several types of speech material (ANSI, 1997). One basic property of the SII is to employ an empirical reference function for each kind of speech material that relates the available speech information (expressed in the SII) to the average recognition rate with the specific speech test material. As long as the tested conditions are closely related to these reference conditions and deviate, e.g. only in the spectral shape of the speech and the stationary background, the SII reaches a very high prediction accuracy (Meyer and Brand, 2013). Hence, the SII is denoted as a reference-based speech intelligibility prediction method whereas the ASR-based method introduced in this paper performs the prediction directly without the need to provide such an empirical reference. Further limitations of the simplifying SII concept have been discussed in several recent studies (Jørgensen and Dau, 2011; Stone et al., 2011, 2012; Hochmuth et al., 2015). The authors argued that better predictions of speech intelligibility can be obtained by taking the temporal modulations of the noise and the speech signal into account.

Cooke (2006) proposed a missing-data ASR-based approach using "glimpses" in noisy speech, which are spectro-temporal regions where a speech signal is least affected by a noise signal, to predict human consonant intelligibility. The ASR system was trained on clean vowel-consonant-vowel logatomes. The noisy speech was then recognized using the glimpses as a mask for the missing-data ASR recognizer, where the glimpses (spectro-temporal regions with positive SNRs) were calculated

using the clean speech and the noise signal and provided ample prior knowledge to the recognizer.

An information theoretic approach which uses an auditory model was presented by Stadler et al. (2007) who refined the idea from Leijon (2002) to use a Hidden Markov Model (HMM) of speech stimuli to estimate sensory information transfer. They estimated the information transmission capability of models of auditory representations of speech in noise and derived speech reception thresholds (SRTs) for a matrix sentence test from the estimated transfer rates. This approach is related to using a HMM based speech recognizer to predict speech intelligibility, but differs in two important aspects from the approach proposed in this study. Firstly, Stadler et al. (2007) generated and analysed models of noisy speech, but these were not tested with noisy speech signals, hence no recognition of noisy speech is actually performed. Secondly, the model of noisy speech was not learned from noisy speech signals but generated from separate models of clean speech and pure noise.

Another auditory-motivated model for human speech recognition mimicking the signal processing that is performed in the elementary auditory parts was proposed by Jürgens and Brand (2009). This model was based on the concept of Holube and Kollmeier (1996) and used an automatic speech recognizer with an auditory model as a front-end and a dynamic-time-warp based back-end. Jürgens and Brand (2009) showed that their model was capable of discriminating CVC and VCV logatomes in noise almost as well as human listeners if the recognizer has perfect prior knowledge. Here, perfect prior knowledge means that the training and test speech material is identical (or "frozen") and the clean speech signal of the to-be-recognized sentence is known to the recognizer. Another downside of this approach was that all parts of the speech signal are assumed to be useful for speech intelligibility, which is not always the case. In reverberant conditions, for example, the early part of the reverberation aids the recognition process while the late part is harmful (Lochner and Burger, 1964; Bradley et al., 2003; Warzybok et al., 2013). Another example is nonlinearly processed speech signals with artifacts, which may result in degraded intelligibility (e.g., Ludvigsen, 1993; Hohmann and Kollmeier, 1995; Rhebergen et al., 2009).

In yet another approach, Jørgensen et al. (2013) estimated the envelope power SNR which takes temporal amplitude modulations into account in order to successfully predict the intelligibility of unprocessed and

processed (reverberation and spectral subtraction) speech in stationary and fluctuating noise conditions. However, this model relied on the concept of an ideal observer implying that a theoretical observer had a perfect prior knowledge about the ("frozen") speech and noise signals and it required the fitting of parameters to an empirical reference condition in order to predict SRTs defined as the SNR yielding 50% intelligibility.

To overcome the limitations and shortcomings of the model approaches listed so far, in the current work we used an ordinary ASR system which was trained and tested using only noisy speech signals on a broad range of SNRs and exploited the fixed semantic structure of the matrix sentence recognition test to obtain an ASR performance that matches human performance. From the noisy sentences, the recognizer could learn during its training procedure which portions of the recordings carried speech information and how reliable that information is in different contexts, i.e. in different words and at different SNRs. Reference-free, objective SRTs were obtained directly from the measured recognition performance of the ASR system, which constitutes an important difference to the SII-based approach which requires defining a reference condition to transfer SII-values to SRTs. Hence, with the SII only differences relative to a reference condition can be predicted, but no reference-free, objective measurements can be performed. On the contrary, the proposed ASR approach predicts reference-free SRTs based on the noisy speech material without any calibration of the system to the empirical data. Strictly speaking, the SRTs are not only predicted, but the very same task that human listeners perform is performed by a standard ASR-based computer model. The lowest resulting SRT is reported across all possible training conditions within the limits of the matrix sentence recognition test materials. Hence, the assumptions about the recognition performance by humans (included either in the ideal observer or in the SNR-to-speech intelligibility relation utilized by the models outlined above) is replaced by a procedural simulation of the task that human listeners perform. The intention of this approach is that the resulting objectively predicted thresholds should be essentially constrained by the testing data and the signal representation and should be independent from the details of the ASR implementation and the training procedure.

In addition, only very few assumptions about the internal representation of the acoustical input signal were made by the standard ASR front-end. It incorporated only basic auditory principles, i.e. time-frequency analy-

sis using a Mel-frequency scaled filterbank, log transform and cepstral analysis by considering only the low-frequency cepstral coefficients that carry vocal tract information and discarding voice-quality carrying higher frequencies. The temporal change of these Mel-frequency cepstral coefficient (MFCC) speech features (i.e. delta and double-deltas) were also included to represent temporal integration and temporal changes within each frequency band. Hence, the extraction of MFCCs and the ASR was generally less complex and included much fewer assumptions about human auditory processing than the auditory models used by Stadler et al. (2007); Jürgens and Brand (2009); Jørgensen et al. (2013).

The ASR system was trained and tested in the same conditions as described in the study of Hochmuth et al. (2015) in which speech intelligibility was measured with listeners with normal hearing using the German matrix sentence test in different noise conditions. Consequently, the ASR-based predictions of this data set were compared to SII-based predictions from Hochmuth et al. (2015) in terms of Pearson s correlation coefficient, bias, and root-mean-square (RMS) errors. Hochmuth et al. (2015) showed that the SII was not able to predict the measured data in the stationary noise conditions as they found no statistically significant correlations of measured and predicted data. This data set therefore constitutes a good benchmark for any new speech intelligibility prediction method and was used within this study to evaluate the proposed ASR-based speech intelligibility model.

6.2 METHODS

6.2.1 Speech intelligibility measurements

The empirical data and the SII-based predictions were taken from Hochmuth et al. (2015). There, a detailed description of the experimental method and setup was provided. Briefly, a series of speech intelligibility experiments in nine different noise conditions was conducted. In the current study, a subset of those six stationary noise conditions and a babble noise condition was considered for which SII-based predictions were available. Note that even though variations of time-dependent, extended SII-estimates exist (see Meyer and Brand (2013) for a review), no valid standardized SII prediction for non-stationary noises exists that could be employed and reported in the study by Hochmuth et al. (2015) in comparison to the empirical data. Hence, the two conditions with

modulated noise were not included here. Speech intelligibility was measured adaptively with listeners with normal hearing with the German matrix sentence test (Wagener et al., 1999a,b,c) to obtain the SRTs. The matrix-type sentences of a fixed syntactical structure were generated from a 50-word base matrix consisting of 10 names, 10 verbs, 10 numerals, 10 adjectives, and 10 nouns. The noises included the stationary test-specific noises of the German, Spanish, Russian, and Polish matrix test (Hochmuth et al., 2012; Warzybok et al., 2015a; Ozimek et al., 2010), the stationary, speech-shaped ICRA1 noise with male and female speaker characteristics (Dreschler et al., 2001), and a multitalker babble noise composed of the recordings of 12 female and 8 male speakers reading different English passages (Auditec, 2006, CD "CD101RW2"). Ten listeners with normal-hearing (pure-tone threshold did not exceed 20 decibels (dB) HL for all octave frequencies from 125 Hz to 8000 Hz) participated in the measurements. The speech signals were presented monaurally to the listener's preferred ear over headphones (Sennheiser HDA200). The adaptive procedure (A1) of Brand and Kollmeier (2002) was used to determine the SRT, where the noise level was fixed at 65 dB SPL and the speech level was varied adaptively to converge to the SRT, while the step size decreased exponentially after each reversal. The SRT was estimated from the psychometric function which was fitted to the data using the maximum-likelihood method. The psychometric function was represented by the logistic function. The order of the measurement conditions was randomized across listeners. The listener's task was to indicate on a touch screen which words she/he understood from a matrix containing all 50 words of the test.

6.2.2 Automatic speech recognizer

The word recognition rate of an ASR system was obtained on the noisy matrix sentences as a function of the training SNR and the testing SNR. An ordinary ASR setup, using the Hidden Markov Toolkit (HTK, Young et al., 2006), was employed, which used Mel-frequency cepstral coefficients (MFCCs, Davis and Mermelstein, 1980) as a front-end, and Hidden Markov Models (HMMs) as a back-end. The front-end transformed a signal waveform into a representation that facilitates the recognition of speech which is also referred to as features. The back-end learned acoustical models of the words to be recognized based on these features,

which were then combined to sentence models using the German matrix sentence grammar.

6.2.2.1 Front-end

As the front-end, MFCCs together with their first and second order temporal derivatives were used forming a feature set that is widely used in ASR. The extraction of MFCC features is usually based on a logarithmically (log) scaled Mel-spectrogram (cf. ETSI, 2003, Standard 201 108). In this study, the log Mel-spectrogram was calculated as follows: First, an amplitude spectrogram using a window length of 25 ms and a window shift of 10 ms was calculated from the input waveform. Then, the linear frequency axis of the amplitude spectrogram was transformed into a Mel-frequency axis by combining the frequency bins from 64 to 8000 Hz with triangular filters into 31 equally-spaced Mel-bands. Finally, the amplitude values were compressed with the decade logarithm. MFCCs were extracted from the log Mel-spectrogram by spectrally processing it with a discrete cosine transform (DCT), where only the first 18 DCT coefficients, which account for spectral modulation frequencies from 0 to 0.29 cycles/Mel-band, were used. The 18 MFCCs were concatenated with their first and second discrete temporal derivatives, which were calculated by applying a temporal slope filter of 5 frames (= 50 ms) length once or twice, respectively. The resulting MFCC feature vectors, which included both derivatives, were 54-dimensional. As stated in the introduction, the log Mel-spectrogram already incorporates some auditory processing principles such as an auditory frequency scale with a limited frequency selectivity and the compressive, logarithmic perception of sound intensity. The subsequent transformation into the cepstrum has no auditory processing background, but rather helps to separate the speaker-specific higher cepstral components from those lower cepstral components relevant for speech recognition. The extraction across five time frames of delta and double-delta MFCC features corresponded to an "effective" amplitude modulation bandpass filter for those modulation frequencies most relevant for speech perception, with center frequencies of about 14 Hz and −3 dB widths of 14 Hz and 10 Hz, respectively.

6.2.2.2 Back-end

On the back-end side, HMMs were used to model speech with whole-word models based on the acoustical representation provided by the front-end. Therefore, the freely available HTK was employed to learn 50 whole-word models, one for each of the 50 words of the base matrix. Each word was modeled with a left-to-right HMM with 16 states. It was assumed that any word of the German matrix test can be represented by a sequence of a maximum of 16 different states, and that the feature vector values for each state were sufficiently well described by a single Gaussian distribution. In other words, if the feature vectors had been spectral energy distributions, a word would have been represented by a sequence of 16 spectral shapes and the energy values would have been assumed to be normally distributed. The number was limited to 16 because states were not allowed to be skipped and hence, the minimum length of any modeled word was 16 frames (= 160 ms). The motivation of using the same number of states for each word was to keep the model simple by making as few assumptions as possible. The model parameters were estimated (learned) in a total of eight iterations on whole matrix sentences, i.e. the speech material was not segmented into single words. In addition to the word models, a silence model, a start and a stop model were trained with four states each. Like the word models, the start and the stop model were left-to-right models where no state was allowed to be skipped, while the silence model allowed transition between all states. The start and stop models modeled possible border effects at the beginning and at the end of a recording, such as the static onset at the beginning of a recording, preventing the silence model from modeling them. The silence model surrounds the sentence which allowed the sentence to be preceded and followed by noise. It also modeled the uncertainty of the listener about when the sentence starts or stops. For the sentence recognition, a language grammar was used to exactly model the German matrix sentence test syntax and hence, the number of available alternatives per word group. The grammar was converted to a word network and used to limit the recognizer to search only for transcriptions with valid matrix syntax.

6.2.3 Predicting SRTs with the automatic speech recognizer

An SNR range from −24 dB to +6 dB was considered, which exceeded the range of expected SRTs from the literature (Hochmuth et al., 2015). Within the considered SNR range, training and testing data sets were generated in 3-dB steps by mixing 120 matrix sentences with portions of the noise signal. The 120 matrix sentences were the same as used in the intelligibility measurements with humans in Wagener et al. (1999b) and Hochmuth et al. (2015). For every mixture of a speech and a noise signal an independent random part of the noise signal was selected, i.e. the noise was not "frozen". For each word of the base-matrix several different recordings were included in the speech material, i.e. the speech was also not "frozen". Hence, the training and the testing data sets may have contained the same portions of the noise signal but were very unlikely to share identical mixtures, just like human listeners were unlikely to be presented with the same stimulus twice.

For the training data set, each of the 120 matrix test sentences was mixed eight times with a randomly chosen portion of the noise signal at each considered SNR. Since within 120 sentences each word occurred exactly twelve times, 96 samples per word-model were present. This guaranteed, that for each model state at least 96 data points were available during training. For the testing data set, each of the 120 matrix test sentences was mixed once with a portion of the noise signal at each considered SNR. Since each matrix sentence contained five words, the word recognition performance was evaluated in 600 decisions (120 sentences × 5 words). The uncertainty of the word recognition performance was estimated by assuming 600 independent binary decisions (correct or incorrect).

The ASR system was trained and subsequently tested on all SNRs, in order to obtain a psychometric function which showed the recognition performance as a function of the testing SNR. This resulted in an $11x11$ matrix which contained the average word recognition rate for all combinations of training SNRs and testing SNRs. This matrix is referred to as the "recognition result map" (RRM) in the remainder of the document. An example of a RRM is depicted in Figure 6.1, A, where the average word recognition rates are represented by the gray scale with white color corresponding to a recognition score of 100% and a black color corresponding to a recognition score of 0%. For each training SNR,

a psychometric function of an ASR system can be obtained from the rows of the RRM.

The SRTs depending on the training SNR were determined from the RRM by linear interpolation. The dotted black-white iso-performance line in Figure 6.1 A, marks the interpolated 50% word correct recognition points, and hence the possible SRTs. The uncertainty was propagated from the measured word recognition rates to the interpolated SRTs, assuming normally distributed errors. From the possible SRTs, the lowest one including a 2-sigma security margin was selected as the predicted SRT.

The reason for using this automatically selected "optimal" training condition that yields the lowest possible SRT is simply to minimize its effect on the predicted thresholds. This should warrant that the predicted SNR primarily depends on the properties of the testing data and the signal representation (features), but not on the properties of the training data.

6.2.4 Speech intelligibility index

The SII-based predictions were taken from Hochmuth et al. (2015). There, the predictions were performed using the SIP-toolbox provided by (Fraunhofer Fraunhofer IDMT, 2014) which implements the calculation of the SII according to the (ANSI, 1997, standard). The SII was computed by dividing the long-term speech spectrum and the long-term noise spectrum into 20 third-octave bands and estimating the weighted average of the SNRs. The band-specific SNRs were weighted using the band-importance function for speech in noise for third-octave bands (Table B2, ANSI, 1997). The reference SII was determined to be 0.214 using the test-specific noise (German matrix test noise) condition. The SRT was then predicted by the SNR that yielded the SII-value of the reference condition. Due to this calibration, the SII-based prediction did not differ from the empirical data in the reference condition.

6.3 RESULTS

6.3.1 Empirical data

The measured SRTs from Hochmuth et al. (2015) to which the ASR-based predictions were compared are briefly discussed here. The empirical SRTs

ranged from −10.9 dB in the Spanish matrix test noise to −6.2 dB in the multi-talker babble noise with a mean SRT of -7.9 ± 1.5 dB across the measurement conditions. Among the language-specific matrix test noises, the highest SRT of −7.2 dB was obtained in the German matrix test noise, confirming that the best energetic masking effect is obtained when the noise spectrum matches the speech spectrum. Slightly and not significantly lower SRTs were observed in the Russian, and Polish matrix test noises as well as in the Icra1 female and Icra1 male noise. In the Spanish matrix test noise, the SRT was 3.7 dB lower than in the German matrix test noise. The highest SRT was measured in the babble noise condition suggesting an additional effect, which could not be explained with purely spectral masking.

6.3.2 ASR-based predictions

An example of the recognition result map (RRM), which represents the recognition performance depending on the training and the testing SNR, is depicted in Panel A of Figure 6.1 for the German matrix test in its test-specific noise. Generally, ASR systems that were trained at high SNRs performed well when being tested at high SNRs and ASR systems that were trained at very low SNRs did not perform above chance level. The lowest SRT in the test-specific noise condition was achieved when training at 0 dB SNR. Hence, this training condition was used to predict the performance of human listeners. The corresponding psychometric function (row of the RRM) is shown in Panel B of Figure 6.1.

For training at a SNR of 0 dB, the recognition performance was found to be about 10% words correct (chance level) for SNRs below −12 dB, and 100% words correct for SNRs above 0 dB. The (interpolated) SRT was -7.6 ± 0.1 dB, and the slope of the psychometric function derived at that point was 16.1 ± 1.2 dB/%. Wagener et al. (1999c) measured an SRT of −7.1 dB and a slope of 17.1 dB/% for listeners with normal hearing in this condition, which coincides remarkably well.

Table 6.1 reports the predicted SRTs along with the empirical data, the SII-based predictions from Hochmuth et al. (2015) and the differences between empirical data and model data in the considered noise conditions.

The absolute deviations from the empirical SRTs in the different noise conditions were consistently smaller for the ASR-based predictions than for the SII-based predictions. In other words, the ASR-based predictions

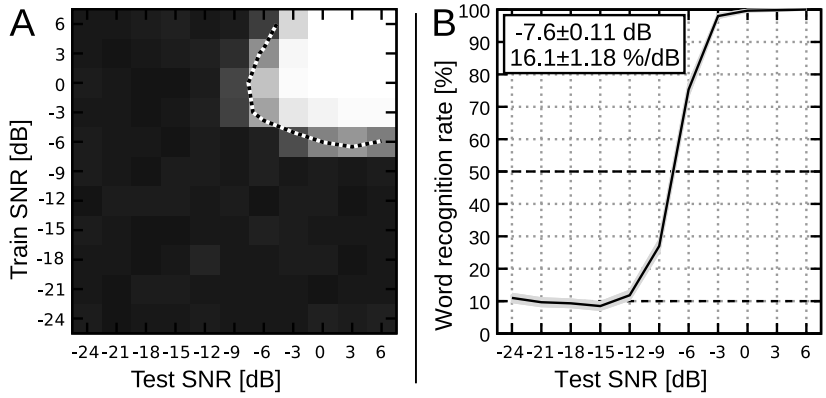

Figure 6.1: *Panel A: Recognition results map (RRM) for the test-specific noise condition (German Matrix noise). Recognition performance of the ASR system depending on the training SNR and the testing SNR in percent word correct is encoded in gray scale, where white corresponds to* 100 *% correct and black to* 0 *% correct. The dashed black-and-white line marks the iso-*50 *%-correct contour. Panel B: The psychometric function obtained with the ASR system trained at* 0 *dB SNR. The recognition performance in percent words correct is plotted over the tested SNRs with the solid lines connecting the data points. The chance level (*10 *%) and the target level (*50 *%) are indicated with dashed lines. The SRT is interpolated linearly between the sampled data points.*

Table 6.1: *Measured SRTs, predicted SRTs, and differences of both in various noise conditions in dB. The employed noises included the stationary test-specific noises for the Matrix sentence test in German (DE), Spanish (ES), Russian (RU), and Polish (PL) language. In addition the male and the female version of the stationary ICRA1 noise was used, and a multitalker babble noise.*

SRTs [dB]	Human	SII		ASR	
System	Measured	Predicted	Diff.	Predicted	Diff.
DE matrix noise	−7.2	−7.2	0.0	−7.6	−0.3
ES matrix noise	−10.9	−6.0	4.9	−13.8	−2.9
RU matrix noise	−8.3	−3.1	5.2	−10.2	−2.0
PL matrix noise	−7.6	−5.5	2.1	−7.6	0.0
ICRA1 male	−7.4	−5.6	1.8	−7.1	0.4
ICRA1 female	−8.0	−5.0	3.0	−7.8	0.1
Multitalker babble	−6.2	−5.1	1.1	−5.5	0.7

were found to be generally closer to the empirical data than the SII-based predictions. For the SII-based predictions, the deviation from the empirical data was found to be at least 1 dB in all noise conditions but the reference condition, in which the SII-based prediction matched the measured data by definition. The maximum deviations from the empirical SRTs were found in the Spanish and Russian matrix test noise conditions, in which the human performance is under-estimated by about 5 dB. The mean value of the SII-based predictions was -5.4 ± 1.2 dB.

For the ASR-based predictions, the most pronounced deviations from the empirical data were also found for the Spanish and Russian matrix noise condition, in which the SRTs were over-estimated by about 3 dB and 2 dB, respectively. For the remaining five of the seven noise conditions the predictions did not differ by more than 0.7 dB from the measured data. The average predicted SRT over all noise conditions was -8.5 ± 2.7 dB.

In Figure 6.2, the predicted SRTs are plotted against the measured SRTs in each noise condition for the SII-based (diamonds) and ASR-based (circles) predictions. The dashed and the dash-dotted lines have a slope of unity and were fitted by minimizing the mean-square error to the SII-based and ASR-based predictions, respectively. The distances between these lines and the bisecting line (solid, black line) were reported as the bias of the respective models, and the RMS error of the predictions was

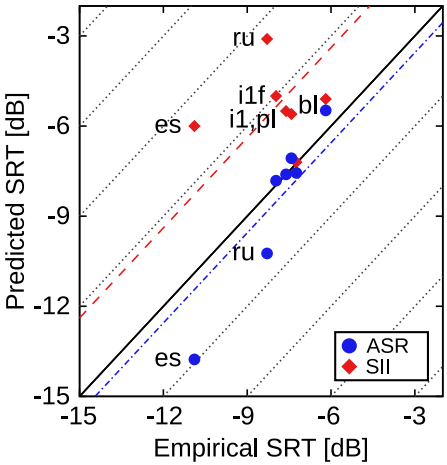

Figure 6.2: *Scatter plot of predicted SRTs against measured SRTs for the ASR system (blue filled circles) and SII (red filled diamonds). The solid line is the bisecting line. All interrupted lines have a slope of unity. The dashed line and the dash-dotted line were fitted by minimizing the mean-square error to the predicted data of the SII-based and ASR-based predictions, respectively. The distance between each line and the bisecting line represents the bias of the predictions. The letters identify the deviating noise conditions where es, ru, and pl refer to the test specific noise of the Spanish, Russan, Polish Matrix sentence test, respectively, i1 refers to ICRA1 (male), i1f refers to ICRA1 (female), and bl to the multitalker babble noise condition.*

calculated from the differences between the predicted and the measured SRTs. In addition, Pearson's correlation coefficients (R^2) between the measured SRTs and the predicted SRTs were calculated along with the probability (significance level $p = 0.05$) of the null-hypothesis (no correlation). Table 6.2 presents the results of the correlation analysis (R^2 and p), the RMS error, and the bias of the different systems.

The RMS of the differences of the SII-based predictions and the empirical data over all noise conditions was 3.1 dB, while the RMS difference of the ASR-based predictions and the empirical data was 1.4 dB. The bias of the SII-based predictions was found to be 2.6 dB, which corresponds

Table 6.2: *Statistical analysis of the predicted SRTs. Pearson's correlation coefficients (R^2) and the probability (p) of the null-hypothesis (no correlation) are reported along with the root-mean-square prediction error and the bias for the SII-based and the ASR-based predictions.*

System	R^2	p	RMS error [dB]	Bias [dB]
SII	0.00	0.987	3.13	2.59
ASR	0.95	0.000	1.35	−0.56

to an under-estimation of the human performance, while the bias of the ASR-based predictions was found to be −0.6 dB, which corresponds to a slight over-estimation of the human performance. No significant ($p = 0.987$) correlation was found between the empirical data and the SII-based predictions, while the predictions of the proposed ASR-based method were found to be highly and significantly correlated ($R^2 = 0.95$, $p < 0.001$) with the empirical data.

6.4 DISCUSSION

The proposed ASR-based method was shown to accurately predict speech reception thresholds (SRTs) in various stationary noise conditions and a babble noise condition for the German Matrix sentence test, whereas the SII-based approach failed to accurately predict them although it was specifically designed for speech intelligibility predictions in stationary noises (cf. Table 6.2). This is remarkable because the SII considers the long-term spectrum of both speech and noise and had been shown to explain those aspects of human speech perception that are primarily due to SNR differences in the frequency domain. The experimental data considered here was well suited for testing this property since the same target speech material was used and the background noises differed primarily in their spectral content.

With the ASR system, on the other hand, it was possible to obtain plausible psychometric functions, like the one depicted in Figure 6.1 B, of which a reference-free SRT and the slope of the psychometric function at that SRT could be derived. Furthermore, the predicted slope of the psychometric function was remarkably close to the measured slope indicating that the ASR-based speech recognition process captures

relevant properties of human speech perception. This outperforming of the SII-based approach by the new reference-free ASR-based approach was unexpected, because it occurred in those conditions where the SII is expected to predict speech recognition in noise very accurately whereas most of the previously reported ASR-based method were only able to predict human speech recognition with a considerable man-machine gap (e.g., Meyer et al., 2011b). It should also be noted that the aim of the current paper was not to optimize the robustness of the ASR system under consideration against noise. Instead, the training and testing of the current system should resemble as much as possible human performance during the matrix sentence recognition test.

There are several reasons why the SII-based approach was not able to predict the measured data in Hochmuth et al. (2015). One important assumption with the SII was that temporal information can be neglected for speech intelligibility in stationary noise. However, Stone et al. (2011, 2012) argued that even stationary noises exhibit temporal amplitude modulation patterns, which may influence speech intelligibility. Also, the band-importance functions that were employed may not have been adequate for the speech material: First, the SII was not designed to predict speech intelligibility measured with speech tests consisting of only very limited speech items (50 words). Furthermore, the standardized band-importance functions were established only for American English. However, they may be language- or even speaker-specific (Wong et al., 2007).

The ASR-based approach learned during the training procedure which portions of the recordings carry speech information and their reliability. These portions were allowed to depend on the training SNR, the language, the speaker, or the type of background noise, which reduced the assumptions about them to a minimum. The ASR-based predictions were also different from the long-term spectrum based SII, in that reference-free SRTs were predicted which did not require any calibration of the system to the empirical data.

In distinction to other speech intelligibility prediction models, e.g. from Stadler et al. (2007), Jürgens and Brand (2009), or Jørgensen et al. (2013), no sophisticated auditory model was needed to predict SRTs for the German matrix sentence test in various stationary and a babble noise condition for listeners with normal hearing. The rather basic auditory representation encoded in the MFCCs seemed to be sufficient as long as

stationary or babble type maskers were considered. This indicates that only a few assumptions about auditory processing might be needed to model speech intelligibility in stationary noise conditions.

Also, in contrast to the existing models which require separate clean speech and noise signals or even frozen speech and noise signals to make predictions (Cooke, 2006; Jürgens and Brand, 2009; Jørgensen et al., 2013), the ASR-based model is trained and tested only with noisy speech signals, just as they were presented to a human listener.

The good coincidence between the ASR-based predictions and the measured data indicates that neither the clean speech signal nor frozen speech or noise signals might be needed to accurately predict speech reception thresholds. The remaining assumptions of the proposed ASR system were that a matrix sentence test is used, that the noisy/mixed speech material exists for several SNRs, and that the physical properties of the audio signal encoded in MFCC/delta/double-deltas were sufficient to perform the recognition task. This makes the proposed setup a suitable candidate for reference-free, objective measures in future work because it allows predictions with fewer assumptions than other models and without the need of any prior measurement.

While the proposed approach employs rather controlled and matched training/testing conditions, which is an easy condition in terms of an ASR task, it was unclear if—even being so simple—an ASR system was able to obtain SRTs as low as those achieved from listeners with normal hearing. Following good scientific practices, the ASR setup was reduced to the minimum of complexity that was required to perform the German matrix sentence recognition task. Contrary to the general expectation that ASR systems perform worse than listeners with normal hearing in acoustically challenging conditions, in the proposed setup an ASR system could perform as well as or even better than those with normal hearing. This is mainly due to the very specific conditions employed here like, e.g. training and testing on the same sentences/noise portions that do not reflect the "standard" ASR problem where neither the speech signal to be recognized nor the interfering noise is not known beforehand. Such a restricted training/recognition setup rather resembles the "optimum detector" assumption employed in some of the psychoacoustical perception models outlined above, even though it differs in detail due to employing a HMM-based recognizer rather than a simple correlation detector. We assumed that the predicted thresholds with this setup were optimal in

the sense that they could not be improved further by using any other training data set, whether it be "multi-condition" or "different speaker". For the prediction of speech intelligibility it is desirable that the predicted values only depend on the test data (not on the training data) and the signal representation (here: MFCCs). However, these restrictions in the training/recognition setup could be relaxed step-by step in future research in order to improve current ASR systems, by systematically investigating their weak points when starting with human and machine performance being on a par with each other.

The proposed method is word-independent which would facilitate its application to matrix-type speech materials of different languages and/or speakers in future work. Since the matrix-type test has been developed so far for 14 languages (Kollmeier et al., 2015) it is possible to measure speech intelligibility in a controlled and comparable way across languages. In future research it could be tested if speech intelligibility predictions for these tests using a language-independent ASR-based model are feasible. This opens the possibility to develop a standard, language-independent way of modeling speech intelligibility as a benchmark for matrix sentence tests and each noise employed.

For more challenging conditions, e.g. modulated noises, reverberation, competing talkers, or processed speech, extensions to the proposed setup, such as using robust ASR features, might be required to perform threshold predictions in future work. However, this was beyond the scope of the current work. Nevertheless, for the ASR system it makes no difference if the distortions originate from the noise signal or from the speech signal itself (e.g. reverberation).

Further, it could be evaluated if the ASR-based prediction method could be used to test if a certain representation of speech signals (e.g. from a sophisticated auditory model with or without a simulated reduction in information transmission due to hearing impairment or cochlear implant excitation patterns) provides sufficient information for decent speech-in-noise recognition performance. Hence, the reference-free ASR-based approach suggested here might be applicable to a variety of applications in hearing research, audiology and communication acoustics.

6.5 CONCLUSIONS

The most important findings of this work can be summarized as follows:

— Speech intelligibility for the German matrix sentence test was accurately predicted for listeners with normal hearing using an ordinary ASR system which modeled the speech perception process in stationary noise conditions and a babble noise condition and outperformed standard SII-based predictions.

— Compared to other speech intelligibility models (e.g. the speech intelligibility index), only few assumptions were needed: Speech reception thresholds were predicted without requiring empirical reference values or assuming an a-priori knowledge-driven optimum detector, thus providing a truly objective measure.

— The proposed method was designed to be easily extendable, e.g. for matrix tests in various languages and different conditions or for the integration of models of impaired signal processing into the front-end, and could potentially spark further research.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft SFB TRR 31 “The active auditory system” and the Cluster of Excellence Grant “Hearing4all”.

7 A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception

ABSTRACT

A framework for simulating auditory discrimination experiments, based on an approach from Schädler et al. (2015b) which was originally designed to predict speech recognition thresholds, is extended to also predict psychoacoustic thresholds. The proposed framework is used to assess the suitability of different auditory-inspired feature sets for a range of auditory discrimination experiments that included psychoacoustic as well as speech recognition experiments in noise. The considered experiments were: 2 kHz tone-in-broadband-noise simultaneous masking depending on the tone length, spectral masking with simultaneously presented tone signals and narrow-band noise maskers, German Matrix sentence test reception threshold in stationary and modulated noise. The employed feature sets included: Spectro-temporal Gabor filter bank features, Mel-frequency cepstral coefficients, logarithmically scaled Mel-spectrograms, and the internal representation of the Perception Model from Dau et al. (1997). The proposed framework was successfully employed to simulate all experiments with a common parameter set and obtain objective thresholds with less assumptions compared to traditional modeling approaches. Depending on the feature set, the simulated reference-free thresholds were found to agree with—and hence to predict—empirical data from the literature. Across-frequency processing was found to be crucial to accurately model the lower speech reception threshold in modulated noise conditions than in stationary noise conditions.

This chapter was accepted on 25th of April 2016 for publication as "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception", M. R. Schädler, A. Warzybok, S. D. Ewert, and B. Kollmeier, in the Journal of the Acoustical Society of America.

7.1 INTRODUCTION

Even though robust automatic speech recognition (ASR) systems have been shown to profit from knowledge about the human auditory system (Hermansky, 1990; Tchorz and Kollmeier, 1999; Kleinschmidt and Gelbart, 2002a; Meyer and Kollmeier, 2011a; Schädler et al., 2012a) and – in return – human auditory signal processing models may profit from the framework and rigid theory behind ASR systems (e.g., Holube and Kollmeier, 1996; Stadler et al., 2007; Jürgens and Brand, 2009) both fields of research have traditionally evolved independently of each other. Typically, any exchange between the two is unidirectional in the sense that (modified) auditory signal processing models are considered as front-ends in ASR, but ASR front-ends are not considered as models of human auditory signal processing. Hence, the aim of this study is to revise the traditional idea of "fitting" auditory models "to the task" in favor of finding universally-valid functional models which are able to perform as well as human listeners in a range of auditory recognition tasks. Such an approach should bridge the fields of automatic speech recognition (ASR), human speech recognition (HSR) and psychoacoustics research. Compared to many models of human auditory signal processing, which are tailored to describe and model specific properties of the human auditory system, ASR features are subject to an extensive set of broad, sometimes even contradictory, demands, e.g., sufficient spectral/temporal detail but good generalization over acoustic conditions. These different objectives (descriptive model vs. universally applicable model) are the reason for auditory models usually requiring considerable modification and engineering towards the appropriate ASR framework before they can be employed as front-ends for ASR purposes. From a modeling point of view, ASR features have desirable properties as a result of the selection process that they undergo in ASR experiments: They are the best known compromise between the diverse demands which are made on the signal representation by robust ASR tasks and, even beyond, audio classification tasks. Hence, auditory-inspired robust ASR features are often simpler than the models by which they were inspired because only the indispensable processing steps for solving the ASR task were actually used. In fact, many common ASR features incorporate only basic auditory signal processing principles such as a limited spectral resolution as well as a compressive intensity perception, e.g., all features which are based on logarithmically scaled

Mel-spectrograms (LogMS). It seems legitimate to ask for the "auditory fidelity" of auditory-inspired ASR features, or in other words, if they show those properties of the auditory system by which they were originally inspired. Hence, one of the aims of this paper is to demonstrate how the "auditory fidelity" of signal representations, including traditional models of auditory signal processing, might be tested and established.

To evaluate ASR features and auditory models on a set of speech recognition and psychoacoustic discrimination tasks with varying complexity and to provide an unbiased, fair comparison between different features/models and empirical data, a common simulation framework which is able to obtain reference-free, i.e., without super-human prior knowledge, objective thresholds is highly desirable. Thus, in a first step, such a framework that allows the simulation of simple and complex auditory discrimination experiments (ADE) using ASR features as well as the output of auditory models with a single universal parameter set is investigated.

Traditional modeling approaches employ predefined features of the change in the signal to be detected and are typically based on signal-to-noise ratios (SNRs) only, such as the power-spectrum model (Patterson and Moore, 1986), the Speech Intelligibility Index (ANSI, 1997), the envelope lowpass-filter model (Viemeister, 1979), or the envelope-power spectrum model (EPSM; Ewert and Dau, 2000). The resulting detection threshold corresponds to a predefined feature value which may be formalized, e.g., by the Signal Detection Theory (Green and Swets, 1966). While these models only use long-term features and thus only require statistical representations of signal and noise, some more refined model versions such as the multi-resolution speech-based ESPM (mr-sEPSM; Jørgensen et al., 2013) require reproducible or so-called "frozen" noise to estimate SNRs in short time frames. More sophisticated modeling approaches (Holube and Kollmeier, 1996; Dau et al., 1997; Jepsen et al., 2008; Jürgens and Brand, 2009), perform a pattern match using an "optimal" detector to predict human performance, thus providing an automatic way of finding the appropriate feature(s) to be detected. However, the exact temporal alignment between template and pattern under consideration can only be secured by a "double-ended" approach, i.e., by deriving the template from a prior knowledge of the target signal alone or at a high SNR and a typical representation of the noise. Moreover, this approach is not able to predict plausible thresholds for the outcome of complex auditory discrim-

ination experiments, such as speech intelligibility tests, without requiring an "optimal" detector that possesses super-human prior knowledge, such as, e.g., the exact temporal alignment of the target or masker signals.

As an alternative, the approach presented by Schädler et al. (2015b) relieves the strong assumptions about the fixed temporal structure of the template, and hence about knowledge of the to-be-recognized target or noise signal *prior* to mixing, by assuming a training phase of a Hidden-Markov-model-based automatic speech recognizer (ASR) at a broad range of signal-to-noise ratios. During this training phase, the ASR system learns the task on noisy data, just like human listeners are assumed to do during an adaptation phase. Unlike other approaches and like human listeners, the ASR system then needs to infer the temporal alignment of the target signal from the noisy mixture. This can be denoted as a pseudo-single-ended approach which only relies on the knowledge of a probabilistically controlled succession of certain automatically learned features, which natively allows the use of processed signals (e.g., including the effect of noise reduction). Furthermore, this approach is reference-free, since the predicted thresholds are not dependent on any reference condition which is used by some traditional model approaches to fit detection parameters (such as, e.g., internal noise) to the average human performance.

Therefore, the modeling approach from Schädler et al. (2015b), originally designed to predict the outcome of the German Matrix sentence speech recognition test, was extended to simulate generic ADE and obtain reference-free objective thresholds. Schädler et al. (2015b) successfully predicted the outcome of the German Matrix sentence test for different types of background noise by simulating the experiment using a standard ASR system. They trained and tested the ASR system with noisy matrix sentences on a broad range of SNRs and determined the speech reception threshold (SRT), i.e., the SNR at which the recognition rate is 50% correct. In the current study, this approach was extended to recognize tone-in-masker and only-masker stimuli which allows to simulate classical psychoacoustic detection and discrimination experiments. A set of general purpose back-end parameters was established with the aim of allowing the simulation of different experiments using different signal representations with the same parameter set. The extended framework with the general purpose parameters is referred to as the simulation framework for auditory discrimination experiments (FADE). The goal of

FADE is to provide a general purpose framework to obtain thresholds which were constrained by the task and the signal representation.

FADE was used to simulate basic, psycho-acoustical experiments and more complex Matrix sentence recognition tasks with a range of feature sets (front-ends). On the side of the psycho-acoustical experiments, simultaneous masking thresholds depending on tone duration were included as well as spectral masking thresholds depending on the tone center frequency. On the side of Matrix sentence recognition tests, speech reception thresholds (SRTs) of the German Matrix sentence test were included in a stationary and a fluctuating noise condition. As signal representations, logarithmically scaled Mel-spectrograms (LogMS), standard ASR features, auditory-inspired ASR features, and the output of a traditional "effective" auditory processing model were employed. Mel frequency cepstral coefficient (MFCC) features were used as standard ASR features. The recently proposed Gabor filter bank (GBFB) and separable Gabor filter bank (SGBFB) features, which were shown to improve the robustness of standard MFCC-based ASR systems Schädler et al. (2012a); Schädler and Kollmeier (2015a), encode spectro-temporal modulation patterns of audio signals and were used as auditory-inspired ASR features. The LogMS was also considered as a signal representation because it represents the common basis for MFCC, GBFB, and SGBFB features. The signal representation of the perception model (PEMO) from Dau et al. (1997), referred to as PEMO features, represented the output of a traditional auditory signal processing model. ASR features are usually used with feature vector normalization, such as mean and variance normalization (MVN) (Viikki and Laurila, 1998), while signal representations in auditory models are not. To assess the effect of MVN, LogMS, MFCC, and PEMO features were employed with and without MVN. All considered experiments were simulated using all feature sets and the obtained thresholds were compared to empirical and model data from the literature.

7.2 METHODS

7.2.1 Experiments

The stimuli, the empirical data, and the PEMO model data for the auditory discrimination experiments were taken from the literature (Moore et al., 1998; Derleth and Dau, 2000; Wagener and Brand, 2005; Jepsen

et al., 2008). While the model and empirical data from the literature were measured using adaptive methods, the simulations using FADE were performed using a constant-stimulus method which is explained in detail in Sec. 7.2.3.

7.2.1.1 Simultaneous masking

The stimuli, the empirical data, and the PEMO model data for the tone-in-noise simultaneous masking experiment were taken from Jepsen et al. (2008). There, a 2-kHz tone signal needed to be detected in the presence of a broadband noise masker. The 500-ms Gaussian noise masker was limited to the frequency range from 20 Hz to 5 kHz and included 50-ms raised-cosine ramps. Detection thresholds corresponding to the 70.7%-correct point on the psychometric function were measured for signal duration from 5 to 200 ms which included 2.5-ms raised-cosine ramps.

7.2.1.2 Spectral masking

The stimuli and the empirical data for the tone-in-noise spectral masking experiment were taken from Moore et al. (1998). The signal was a tone and the masker a 80-Hz wide Gaussian noise centered at 1 kHz and presented at 45 dB SPL. Detection thresholds corresponding to the 79.4%-correct point on the psychometric function were measured. The tone frequencies considered in this work were those at which the masking effect was expected to dominate the absolute hearing thresholds: 0.75, 0.90, 1.00, 1.10, 1.25, and 1.50 kHz. The original study considered more conditions including noise signals, tone maskers, additional masker levels, and additional center frequencies. The PEMO model data was taken from Derleth and Dau (2000), which used the same model parameters as Dau et al. (1997). In contrast to the original papers, the thresholds are presented in dB SPL rather than in dB masking. Therefore, the dB masking values were transformed to dB SPL using the absolute hearing thresholds defined in (ISO, 2003, Standard “226: 2003”).

7.2.1.3 German Matrix sentence test

The stimuli and the empirical data for the speech intelligibility experiment were taken from Wagener et al. (1999c) and Wagener and Brand (2005).

In the sentence test from Wagener et al. (1999c), listeners needed to repeat sentences of five words with a fixed syntactical structure which were presented in noise. The SNR which corresponded to the 50%-correct point on the psychometric function, i.e., the SRT, was measured using an adaptive method. The speech material is phonetically balanced and represents the phonetic variety of the German language. In addition to the unmodulated test-specific noise condition, a condition with a single-speaker modulated speech noise from a male speaker at normal level, the IRCA5 noise from Dreschler et al. (2001), was considered. The corresponding empirical thresholds were taken from Wagener and Brand (2005).

7.2.2 Signal representations

The logarithmically scaled Mel-spectrogram (LogMS) is the basis for all considered ASR features (MFCC, GBFB, SGBFB) in this study. Mel-spectrograms were extracted from an amplitude spectrogram of the input waveform with a window length of 25 ms and a window shift of 10 ms. Therefore, the linear frequency axis of the amplitude spectrogram was transformed into a Mel-frequency axis by combining the frequency bins from 64 Hz to 8 kHz with triangular filters into 31 equally-spaced Mel-bands. Finally, the amplitude values are compressed with the decade logarithm. An example of a LogMS is depicted in the upper panel of Fig. 7.1. This 31-dimensional signal representation is referred to as LogMS features.

7.2.2.1 Mel frequency cepstral coefficients

MFCCs are widely used in ASR and acoustic detection tasks and are often used as a baseline. In this work, they were extracted from LogMSs by applying a discrete cosine transform (DCT) in the spectral dimension. Subsequently, the MFCCs corresponding to quefrencies above 0.58 cycles/Mel-band were removed and the remaining 18 MFCCs were concatenated with their first and second order discrete temporal derivative. The temporal derivatives are also called deltas and double deltas and were extracted by applying a slope filter with a total length of 5 frames once or twice respectively. The 54-dimensional MFCC features were used with mean and variance normalization as explained in Sec. 7.2.2.5.

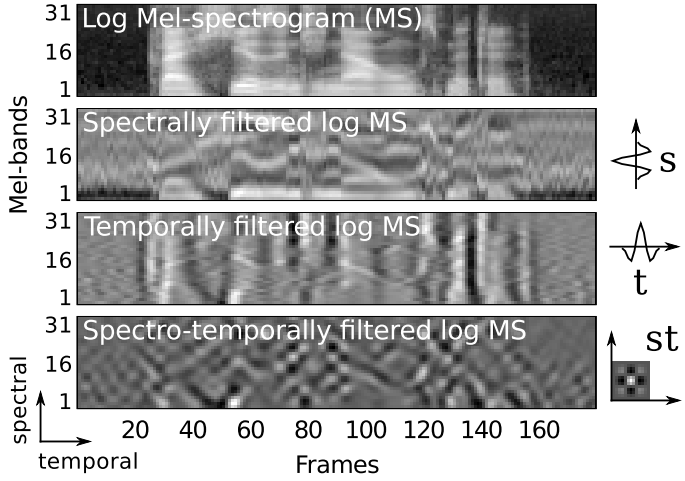

Figure 7.1: *Taken from Schädler and Kollmeier (2015a). The LogMS of clean speech in the upper panel is 2D-convolved with a spectral 1D filter* s*, a temporal 1D filter* t *and the corresponding spectro-temporal 2D filter* st*. The result of the filtering process is depicted to the left of the corresponding filter. The amplitude of the 2D filters and (filtered) spectrograms is encoded in gray scale, where white encodes high amplitude and black encodes low amplitude.*

7.2.2.2 Gabor filter bank features

GBFB features were successfully employed as robust features for ASR by Moritz et al. (2013) as well as robust features for acoustic event detection by Schröder et al. (2013). They are auditory-inspired and extract spectro-temporal modulation patterns from LogMS using 2D Gabor filters. The shapes of the 2D Gabor filter that were used are depicted in Fig. 7.2 and were inspired by patterns found in neural correlates in the auditory cortex of cats by Qiu et al. (2003). To extract GBFB features, the LogMS was 2D convolved with each of the 2D GBFB filters. Each filtered version was subsequently (critically) down-sampled in spectral dimension by a quarter of the width in spectral dimension of the corresponding 2D filter. The filtered and down-sampled versions of the LogMSs were than concatenated and formed a 455-dimensional feature vector. Extensive descriptions of the GBFB feature extraction were given in Schädler et al. (2012a); Schädler and Kollmeier (2015a). GBFB features were used with mean and variance normalization as explained in Sec. 7.2.2.5.

7.2.2.3 Separable Gabor filter bank features

The difference between GBFB and SGBFB features is that SGBFB features are extracted with two separate modulation filter banks, a spectral and a temporal one, instead of using a filter bank of spectro-temporal filters. Nonetheless, they cover the same spectro-temporal modulation space. The SGBFB approach was shown to reduce the complexity of the features and even to improve the robustness of an ASR system (Schädler and Kollmeier, 2015a). All SGBFB filter functions and the corresponding separable 2D filter functions of all combinations of spectral and temporal SGBFB filters are depicted in Fig. 7.3. In the current study 1020-dimensional SGBFB features were extracted using the full set, i.e., all nine spectral and all nine temporal filters, which are referred to as SGBFB features. An extensive description of the SGBFB feature extraction was given in Schädler and Kollmeier (2015a). In addition to the 1020-dimensional SGBFB features, a reduced set of 255-dimensional SGBFB features which does not use the filters that are marked with I (for imaginary phase) in Fig. 7.3, were considered and are referred to as SGFB-RR features. Due to its design, the SGBFB allows to apply only the spectral or only the temporal modulation filtering. A set of features which was extracted using only temporal R (for real phase)

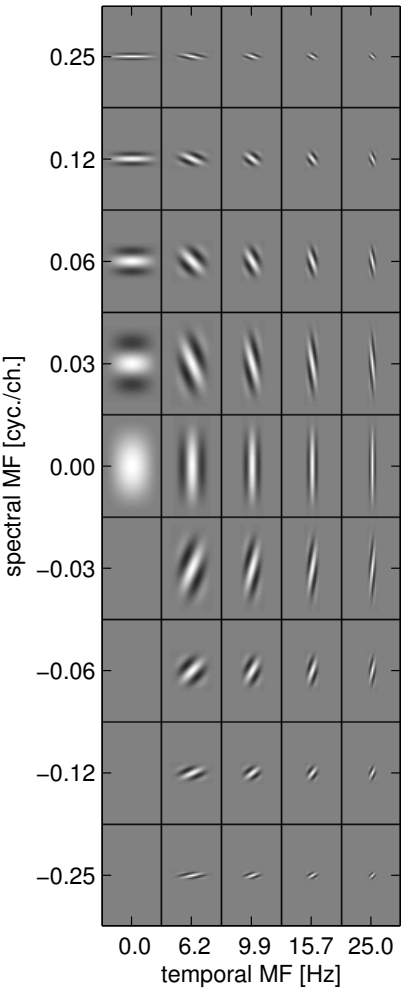

Figure 7.2: *Taken from Schädler et al. (2012a). Filter functions of the 2D Gabor filter bank (GBFB) filters. Each tile represents the filter function of a spectro-temporal 2D Gabor filter, where the horizontal axis within each tile is the temporal one and the vertical axis is the spectral one. They are sorted by their spectral and temporal center modulation frequencies. The amplitude of the 2D filter functions is encoded in gray scale, where white encodes high amplitude and black encodes low amplitude.*

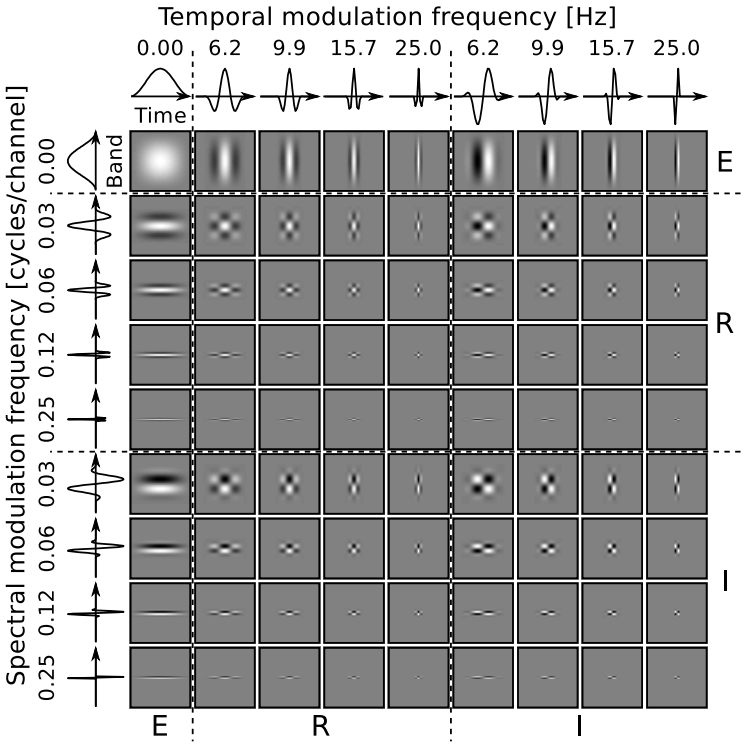

Figure 7.3: *Taken from Schädler and Kollmeier (2015a). All possible combinations of spectral and temporal 1D Gabor filter bank filters, and their equivalent, separable spectro-temporal 2D filter functions. Each tile represents a separable spectro-temporal 2D filter function, with the horizontal axis within each tile being the temporal and the vertical axis being the spectral one. The 1D filters, depicted above and to the left of the 2D filters, are sorted by spectral and temporal center modulation frequencies, and are grouped according to the part of the complex 1D Gabor filter which is used: envelope (E), real (R), imaginary (I). The amplitude of the 2D filters is encoded in gray scale, where white encodes high amplitude and black encodes low amplitude.*

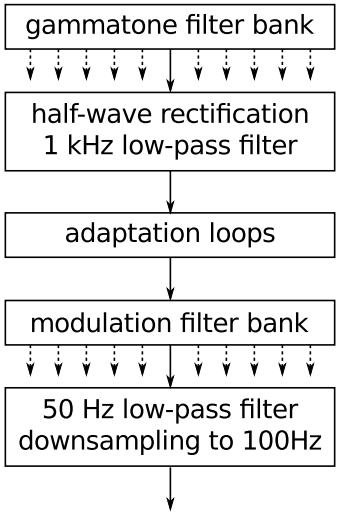

Figure 7.4: *Modified from Dau et al. (1997). Block diagram of the auditory signal processing which is used to calculate the internal representations with PEMO, also referred to as PEMO features. Essentially this is the model from Dau et al. (1997) up to the modulation filter bank. The low-pass filtering at* 50 *Hz and the down-sampling to* 100 *Hz is added to make the internal representation compatible for the use in the recognition system.*

and E (for envelope) filters is referred to as SGBFB-R-T, and another set of features which was extracted using only spectral R and E filters is referred to as SGBFB-R-S. All SGBFB based features were used with mean and variance normalization as explained in Sec. 7.2.2.5.

7.2.2.4 Perception Model

The PEMO was successfully used to model various experiments in psychoacoustics, (e.g., Dau et al., 1997; Verhey et al., 1999; Derleth and Dau, 2000). It was introduced by Dau et al. (1996a,b) and later extended with a temporal modulation filter bank by Dau et al. (1997). The PEMO includes a signal processing part (front-end) which effectively models several aspects of the human auditory system. In the current study the PEMO front-end from Dau et al. (1997) is used to extract features from input waveforms. Therefore, the freely available implementation from Søndergaard and Majdak (2013) at git commit *cc9c0d3c* was used, which considers auditory filters in the frequency range from 80 Hz to 8 kHz and temporal modulation frequencies in the range from 0 Hz to 150 Hz. A block-diagram of the PEMO feature extraction is depicted in Fig. 7.4. A Gammatone filter bank was used to model the response of the basilar membrane to the input signal. The subsequently applied half-wave recti-

fication and the 1 kHz low-pass filter model the hair cell deflection. The adaptation loops account for temporal properties of the nerve cell firing probability at different stages of the auditory pathway. The output of the modulation filter bank was low-pass filtered and down-sampled to 100 Hz. These 275-dimensional feature set is referred to as PEMO features.

7.2.2.5 Feature normalization

Feature vector normalization such as mean and variance normalization (MVN) or histogram equalization were shown to improve the robustness of ASR systems (Viikki and Laurila, 1998; De La Torre et al., 2005) and are usually employed in conjunction with robust ASR features. The auditory models used to explain psycho-acoustical experiments usually do not contain a similar processing step. This is why in the current study by default all ASR features (MFCC, GBFB, and SGBFB) are used with per-utterance/per-stimulus MVN, while the LogMS and the PEMO features are not.

In order to assess the effect of feature normalization, LogMS, PEMO and MFCC features were tested with and without MVN. These feature sets are indicated by the suffix MVN and NOMVN, respectively.

7.2.3 Simulation framework for auditory discrimination experiments

The simulation framework for auditory discrimination experiments (FADE) is based on the approach from Schädler et al. (2015b), where an ASR recognition system was used to simulate—and hence predict the outcome of—the German Matrix sentence test with only few assumptions compared to traditional speech intelligibility prediction models. Here, this approach was extended to simulate tone-in-noise detection (i.e., tone-in-noise from only-noise discrimination) experiments. A reference implementation of FADE is available online[1].

7.2.3.1 Front-end

In the original work by Schädler et al. (2015b), only MFCCs were used as the front-end, while in this work all signal representations presented in Sec. 7.2.2 were employed with the FADE.

[1] URL: http://medi.uni-oldenburg.de/FADE

7.2.3.2 Back-end

The back-end used in FADE is the same as in Schädler et al. (2015b). HTK was used to build left-to-right whole-word/stimulus Hidden Markov Models (HMMs) models with 6 states per word/stimulus and Gaussian Mixture Models (GMMs) with one component per state. For each training condition, which in the case of the German Matrix sentence test is determined by the SNR and for the psychoacoustic experiments by the absolute tone level, the GMM/HMM parameters are estimated (learned) in a total of 8 iterations. Since the material of the German Matrix test consists of 50 words, 50 whole-word models were learned during the training period. For the tone-in-noise detection experiments, two models were trained: A model for the stimuli in which the target is present (tone plus noise) and a reference one for the stimuli in which the target is absent (noise only). In addition to the word/stimuli models, a START, a STOP, a PRE-SILENCE, and a POST-SILENCE model were trained for each training condition. The START/STOP model covers border artifacts which are common to all recordings of a training condition, while the PRE/POST-SILENCE models represent the indistinguishable signal parts before and after the speech/target. All four are shared between all sentences/stimuli of a training condition. The grammar, in HTK-terms, for a sentence/stimulus was: (START PRE-SILENCE $sentence/stimulus POST-SILENCE STOP), where $sentence = ($word1 $word2 $word3 $word4 $word5) and $stimulus = (reference | target). The corresponding grammar was converted to a word network and used to limit the recognizer to search only for transcriptions with valid syntax for the corresponding experiment. This implements the knowledge of a trained listener, who knows about the grammatical structure as well as about the limited vocabulary of the Matrix test. The effect of the number states per model and the number of states per special model (START, STOP, PRE-SILENCE, POST-SILENCE), and the number of training iterations was assessed in Sec. 7.3.4.

7.2.3.3 Simulation

The regions of interest of the values for the independent variables were defined as follows: For the simultaneous masking experiment, tone levels from +45 to +75 dB SPL in 5-dB steps were considered. For the spectral masking experiment, tone levels from −10 to +50 dB SPL in 5-dB steps

were considered. For the German Matrix sentence experiments, SNRs from −24 to +6 dB in 3-dB steps were considered.

For each of these values, datasets for training and testing were generated in the same manner. For the tone-in-noise masking experiments, the two different types of stimuli (target and reference) were generated with random noise, such that a repetition of the same stimulus waveform is practically impossible. For the German Matrix sentence experiments, the 120 available sentences were mixed with the noise signal with random temporal offsets, such that a repetition of the same waveform is practically impossible even if the same sentences were mixed several times with the same noise signal. The 120 sentences contained each word of the 50-word vocabulary exactly twelve times, and mixing all sentences once with random portions of the noise signal resulted in twelve samples per word.

From these (statistical) *pools*, which directly reflect the difficulty of the corresponding recognition task at a given tone level or SNR, a number of samples was drawn and declared as the test data. Because the performance limiting factor, i.e., the difficulty of the task, is inherent to the test data under its projection into the feature space, an optimal training data set was desirable. Hence, the training data sets were drawn from the same *pool* as—but separate from—the test data sets. By this means, we aimed to minimize the influence of the training data set and at the same time to maximize the influence of the test data set on the recognition scores.

The recognition of all 120 available sentences of the German Matrix test produces 600 binary (correct or incorrect) decisions, which was chosen to be the size of the test data sets. It should be noted that each Matrix sentence results in five binary decisions, one for each word, while a presented psychoacoustic stimulus only results in one binary decision. The size of the training data sets of 96 samples for each word/stimulus was assessed in Section 7.2.3.4. For the Matrix sentence test, these were achieved by mixing all sentences eight times with random portions of the noise signal. Features were then extracted from the generated training and test data sets.

For each condition (e.g., speech in fluctuating noise) separately, models were trained and tested for all considered values of the independent variable. For example, 11 models—one for each considered SNR—were trained on speech in fluctuating noise and each subsequently tested in the 11 considered SNR conditions, which resulted in $11 \times 11 = 121$ recognition

scores. These were represented by a (square) matrix called "recognition result map" (RRM), where each row represents a psychometric function of which the value of the independent variable at a given target threshold could be derived. For the Matrix sentence test, the SNR at 50%-correct, which is the standard procedure with human listeners, was determined. For the psychoacoustic experiments, instead of the 50%-correct point on the psychometric function (i.e., the SRT) the corresponding target %-correct point was considered. For each psychometric function, the value of the independent variable at the corresponding target %-correct point was interpolated together with its estimated uncertainty due to the size of test data set. Thus, for the tone-in-noise experiments, several levels at threshold depending on the training level, and for the German Matrix sentence test, several SRTs dependeding on the training SNR were available. As the result of the simulation, the lowest value at threshold was reported, where two standard deviations of margin were considered in order to report the outcome with the lowest 95th percentile (assuming normal distributions). This automatic determination of the optimal training data set, which may depend on the task itself, the amount of training data and the feature representation, is aimed to reduce its influence on the simulation results.

7.2.3.4 General purpose parameter set

At the core of FADE a set of general purpose parameters exists which was employed for all features and experiments, the simplest task being the detection of a tone, the most difficult the discrimination of words in modulated noise, the lowest-dimensional features being 31-dimensional LogMS features and the highest-dimensional being the 1020-dimensional SGBFB features. These parameters were:

— HMM states START/STOP: 6

— HMM states per model: 6

— Training samples per model: 96

— Training iterations: 8

These parameters were considered to be especially important when differently complex features and tasks are involved. To demonstrate that the chosen parameter values are optimal up to ±1 dB for different features

in differently complex experiments, a set of features and experiments was performed when varying the parameter values. Optimal here means that the systems obtained the highest possible recognition rates which translates to the lowest possible thresholds, an optimization scheme commonly used in the field of ASR. Optimal here does not mean that the results were close to the empirical results, which is an optimization scheme commonly used in the field of psychoacoustic modelling. The considered values were:

— HMM states START/STOP: $1, 2, 3, 4, \mathbf{6}, 8, 12, 16, 24$

— HMM states per model: $1, 2, 3, 4, \mathbf{6}, 8, 12, 16, 24$

— Training samples per model: $12, 24, 48, \mathbf{96}, 192, 384$

— Training iterations: $1, 2, 3, 4, 6, \mathbf{8}, 12, 16, 24$

Each of the parameters was varied while the others were left unchanged. Simulations of the simultaneous masking experiment and the German Matrix sentence test in the test-specific noise condition were performed with varied parameter values using MFCC and PEMO features.

7.2.3.5 Uncertainty calculation

The uncertainty of the simulated outcomes due to the limited test data, which was 600 binary decisions per condition, was estimated using bootstrapping. It turned out to be about 2.1 percentage points (pp) at 50% correct, about 1.8 pp at 75% correct, and about 1.2 pp at 90% correct. These estimated uncertainties were assumed to be normally distributed and propagated to derived values, such as SRTs or thresholds, where possible. The uncertainty due to the limited test data was not assessed as it would have required re-running the training stage several times with different data. In addition, the limited step size of training and test conditions could present another source of uncertainty, which was not assessed either. Hence, the uncertainties reported here only include those due to the limited test data and should be considered orientative. Nonetheless, the uncertainty can be assumed to be about 1 dB, which was justified in Sec. 7.2.3.4.

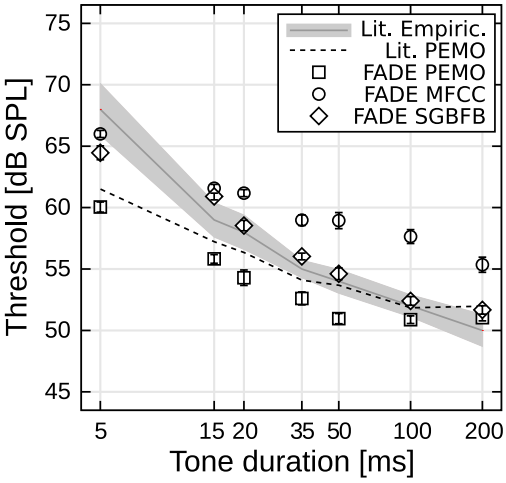

Figure 7.5: *Simulated detection thresholds for the simultaneous masking experiment depending on the tone duration with PEMO, MFCC, and SGBFB features alongside the empirical data and PEMO data from the literature Jepsen et al. (2008). The gray area indicates the 1-sigma uncertainty of the empirical data.*

7.3 RESULTS

Apart from the parameter variation experiment, all simulations were performed with all features. The results are presented in tables and selected results are additionally plotted.

7.3.1 Simultaneous masking

Figure 7.5 depicts the simulated detection thresholds depending on the tone duration with PEMO, MFCC and SGBFB features alongside the empirical results and PEMO model results from the literature (Jepsen et al., 2008). Table 7.1 reports the corresponding results in numerical form for all considered feature sets, and in addition, the average detection threshold over all conditions.

FADE was able to predict detection thresholds for a simultaneous masking experiment with a variety of front-ends. All simulated thresholds were consistent with the empirical thresholds within $\pm 10\,\mathrm{dB}$, i.e., in the correct order of magnitude. MFCC features resulted in the most pronounced over-estimation of the empirical thresholds, with an average detection threshold of $60.0 \pm 0.2\,\mathrm{dB}$ SPL and PEMO features resulted in the most pronounced under-estimation of the empirical thresholds with an average detection threshold of $53.7 \pm 0.2\,\mathrm{dB}$ SPL, while the

Table 7.1: *Simulated detection thresholds in dB SPL for the simultaneous masking experiment depending on the tone duration and the feature set. The empirical data was taken from Jepsen et al. (2008).*

Features	Tone duration							Average
	5 ms	10 ms	15 ms	35 ms	50 ms	100 ms	200 ms	
Empirical	68.0±2.1	59.0±1.4	58.0±1.4	55.0±0.7	54.0±1.0	52.0±0.9	50.0±1.3	56.6±0.5
LogMS	63.0±0.3	60.6±0.7	58.4±0.5	56.9±0.2	54.0±0.5	50.7±0.3	48.2±0.5	56.0±0.2
LogMS-MVN	63.7±0.4	61.6±0.2	61.5±0.2	56.8±0.3	57.5±0.3	52.5±0.5	53.1±0.5	58.1±0.1
MFCC	66.0±0.3	61.6±0.3	61.2±0.3	59.0±0.4	58.9±0.7	57.6±0.6	55.4±0.6	60.0±0.2
MFCC-NOMVN	66.0±0.3	61.3±0.3	61.3±0.3	58.5±0.4	58.3±0.5	56.2±0.4	53.9±0.5	59.4±0.2
GBFB	64.8±0.5	60.8±0.3	59.0±0.5	56.5±0.3	56.1±0.3	52.6±0.4	51.9±0.5	57.4±0.2
SGBFB	64.5±0.6	60.9±0.3	58.5±0.4	56.0±0.3	54.6±0.5	52.4±0.4	51.7±0.3	56.9±0.2
SGBFB-RR	65.5±0.3	61.2±0.3	60.4±0.4	56.8±0.3	55.8±0.4	52.1±0.4	51.9±0.4	57.7±0.1
SGBFB-R-S	64.3±0.5	61.8±0.2	61.4±0.2	57.5±0.5	56.7±0.3	52.7±0.3	52.5±0.6	58.1±0.2
SGBFB-R-T	65.1±0.5	60.5±0.3	60.1±0.4	56.0±0.3	55.9±0.4	54.0±0.6	52.1±0.6	57.7±0.2
PEMO	60.0±0.4	55.8±0.3	54.3±0.6	52.6±0.5	51.0±0.4	50.9±0.3	51.0±0.3	53.7±0.2
PEMO-MVN	60.2±0.4	56.0±0.4	54.5±0.6	52.8±0.5	51.2±0.4	51.2±0.3	51.9±0.3	54.0±0.2

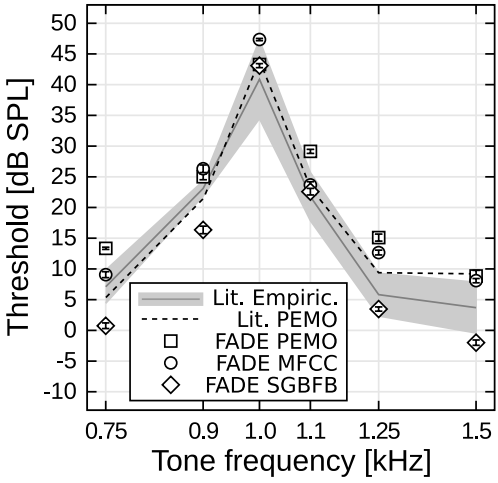

Figure 7.6: *Simulated detection thresholds for the spectral masking experiment depending on the tone center frequency in Hz with PEMO, MFCC, and SGBFB features alongside the empirical data and PEMO data from the literature (Moore et al., 1998; Derleth and Dau, 2000). The gray area indicates the 1-sigma uncertainty of the empirical data.*

empirical results showed an average detection level of 56.6 ± 0.5 dB SPL. Simulation results with all other features lay between simulation results with MFCC and PEMO features. The simulated thresholds with GBFB based features (GBFB, SGBFB, SGBFB-RR, SGBFB-R-S, SGBFB-R-T) were consistently found to be close to the empirical thresholds.

The simulated thresholds with PEMO features are generally about 2 dB lower than the PEMO data from the literature, over-estimating the empirical thresholds for tone-durations shorter than 100 ms. The simulated thresholds with MFCC features under-estimate the empirical thresholds for tone-durations longer than 15 ms. The simulated thresholds with SGBFB feature resemble the empirical thresholds remarkably well. Deviations of simulated thresholds from the empirical data are further analyzed in Sec. 7.3.5.

7.3.2 Spectral masking

Figure 7.6 depicts the simulated detection thresholds depending on the tone center frequency in Hz with PEMO, MFCC and SGBFB features alongside the empirical results and PEMO model results from the literature (Moore et al., 1998; Derleth and Dau, 2000). Table 7.2 reports the corresponding results in numerical form for all considered feature sets, and in addition, the calculated 20-dB-bandwidth.

Table 7.2: *Simulated detection thresholds in dB SPL for the spectral masking experiment depending on the tone center frequency in Hz and the feature set. The full widths were calculated at −20 dB from the data. The empirical data was taken from Moore et al. (1998).*

Features	Tone center frequency						Width [Hz]
	750 Hz	900 Hz	1000 Hz	1100 Hz	1250 Hz	1500 Hz	
Empirical	7.1±2.7	23.1±1.2	40.9±6.4	21.8±3.9	5.8±3.5	3.7±4.1	229.5±38.9
LogMS	3.8±0.2	15.6±0.4	42.1±0.3	24.6±0.4	2.1±0.5	-1.1±0.4	192.3±2.9
LogMS-MVN	5.1±1.0	24.4±0.8	47.3±0.2	42.6±0.4	4.2±0.5	2.2±0.6	246.6±3.3
MFCC	9.1±0.5	26.3±0.6	47.4±0.2	23.7±0.5	12.7±0.4	8.1±0.3	179.6±3.4
MFCC-NOMVN	1.3±0.3	16.6±0.3	43.0±0.3	21.5±0.4	2.8±0.3	-2.7±0.4	168.7±2.1
GBFB	2.4±0.4	18.6±0.6	43.3±0.3	23.2±0.4	5.4±0.6	-1.2±0.4	180.4±3.0
SGBFB	0.8±0.5	16.4±0.6	43.1±0.3	22.6±0.5	3.5±0.3	-2.0±0.4	172.3±3.3
SGBFB-RR	1.7±0.4	16.4±0.5	43.0±0.4	21.6±0.6	3.4±0.4	-1.9±0.4	168.3±3.3
SGBFB-R-S	5.1±0.9	21.0±0.8	43.2±0.4	29.9±0.5	4.5±0.6	-1.3±0.5	229.4±4.7
SGBFB-R-T	4.2±0.9	22.0±0.6	47.2±0.2	28.2±0.7	7.2±1.1	2.7±0.7	186.0±5.2
PEMO	13.3±0.2	25.0±0.5	43.2±0.3	29.1±0.3	15.1±0.6	8.8±0.9	284.8±7.3
PEMO-MVN	16.5±0.2	35.7±0.3	44.6±0.4	37.4±0.5	19.9±0.7	13.2±0.2	396.2±5.8

FADE was able to predict detection thresholds for the spectral masking experiment with a variety of front-ends. Almost all simulated thresholds are within ±10 dB of the empirical thresholds, i.e., in the correct order of magnitude. Only the PEMO and LogMS features with MVN resulted in thresholds outside that range. Generally, the simulations with all features show the highest thresholds at the noise center frequency of 1000 Hz and decrease as the tone frequency increases or decreases. Consistent with the results from the simultaneous masking experiment, the simulated thresholds with MFCC features exhibit the highest on-masker (1000 Hz) thresholds with 47.4 ± 0.2 dB SPL and the simulated thresholds with PEMO features, with 43.2 ± 0.3 dB SPL, one of the lower thresholds. These are—unlike in the simultaneous masking experiment—higher than the empirical threshold, which was 40.9 ± 6.4 dB SPL. The empirical 20-dB-bandwidth was calculated to be 229.5 ± 38.9 Hz. Almost all simulated results fell into the 2-sigma range (151.3 to 306.7 Hz) and hence did not differ significantly from the empirically derived bandwidth. Only the PEMO features with MVN exceeded this range with a bandwidth of 396.2 ± 5.8 Hz. All ASR features (MFCC, GBFB, and SGBFB) result in rather narrow bandwidths around 180 Hz, e.g., using SGBFB features, 172.3 ± 3.3 Hz.

The simulated thresholds with PEMO features were found to be similar or higher than the PEMO model data reported by Derleth and Dau (2000). With MFCC features, the simulated thresholds resembled the empirical data well while with SGBFB features the thresholds on the low frequency flank were over-estimated by about 6 dB. Deviations of simulated thresholds from the empirical data are further analyzed in Sec. 7.3.5.

7.3.3 German Matrix sentence test

The recognition result map (RMM), which is the matrix that contains the recognition rates depending on the training and the test condition, and its evaluation is illustrated in Fig. 7.7 for the simulation results of the German Matrix sentence test with MFCC features. In Panel A, the RRM, i.e., the recognition performance depending on the training and test SNR, is depicted in gray-scale, where black corresponds to 0%-correct and white to 100%-correct. The iso-50%-correct contour is indicated by the dotted black-and-white line and the lowest achievable SRT, which at the same time is the simulation result, is indicated by

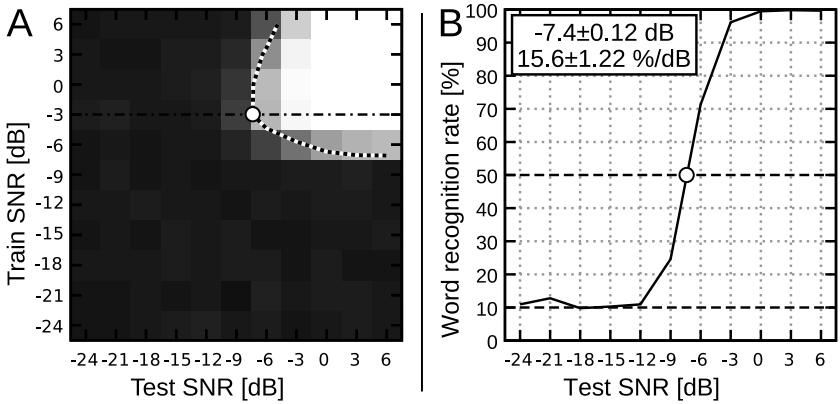

Figure 7.7: *Panel A: Recognition result map (RRM) for the test-specific noise condition with MFCC features. The obtained recognition performance is plotted depending on the training and testing SNR. The word recognition rates are encoded in gray-scale, with white representing* 100% *correct and black* 0% *correct. The dotted black-and-white line marks the iso-*50%*-correct contour. The dash-dotted line marks the training SNR which resulted in the lowest achievable test SNR at* 50%*-correct WRR (SRT). The white circle indicates the predicted SRT. Panel B: Word recognition rates depending on the test SNR for the system that achieves the lowest SRT (cf. the dash-dotted line in Panel A). The chance level (*10%*) and the* 50%*-threshold are marked with dashed lines. The white circle indicates the simulated SRT. The box shows the estimated SRT and slope of the psychometric function, respectively.*

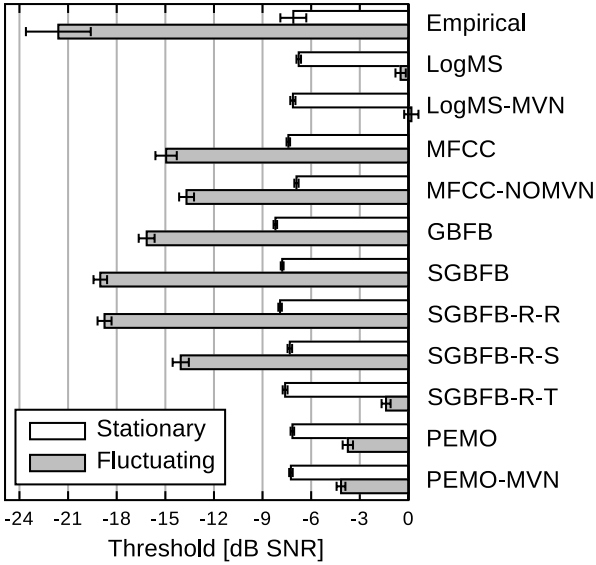

Figure 7.8: *Simulated SRTs in dB SNR for the German Matrix sentence test depending on the noise condition and the feature set. The empirical data was taken from Wagener et al. (1999c) (stationary) and Wagener and Brand (2005) (fluctuating).*

a circle. The corresponding training condition (−3 dB SNR) is marked with a dash-dotted line and the corresponding psychometric function is depicted in Panel B. As expected, the recognition results are at chance level (10%-correct) at low SNRs and tend towards 100%-correct for high SNRs.

Figure 7.8 depicts the simulated SRTs depending on the noise condition and the employed feature set alongside the empirical results from the literature (Wagener et al., 1999c; Wagener and Brand, 2005). Table 7.3 reports the corresponding results in numerical form for all considered feature sets and, in addition, the effect of modulation which is reported as the difference of the SRT in the modulated noise condition (ICRA5) and the test-specific noise condition (Olnoise).

For the stationary noise condition, the simulated SRTs were found to be in the range from −8.2 to −6.7 dB SNR, where the empirical value

Table 7.3: *Simulated SRTs for the German Matrix sentence test depending on the noise condition and the feature set in dB SNR. The empirical data was taken from Wagener et al. (1999c) (Olnoise) and Wagener and Brand (2005) (ICRA5). The effect of modulation is reported as the difference of the SRT in the modulated noise condition (ICRA5) and the test-specific noise condition (Olnoise).*

System	Olnoise SRT [dB]	ICRA5 SRT [dB]	Modulation effect [dB]
Empirical	-7.1±0.8	-21.6±2.0	-14.5±2.2
LogMS	-6.8±0.1	-0.5±0.3	+6.3±0.4
LogMS-MVN	-7.1±0.2	+0.2±0.4	+7.3±0.5
MFCC	-7.4±0.1	-15.0±0.7	-7.5±0.7
MFCC-NOMVN	-6.9±0.1	-13.7±0.5	-6.8±0.5
GBFB	-8.2±0.1	-16.2±0.5	-7.9±0.5
SGBFB	-7.8±0.1	-19.0±0.4	-11.2±0.4
SGBFB-RR	-7.9±0.1	-18.8±0.4	-10.8±0.4
SGBFB-R-S	-7.3±0.1	-14.1±0.5	-6.7±0.5
SGBFB-R-T	-7.6±0.2	-1.4±0.3	+6.2±0.3
PEMO	-7.2±0.1	-3.7±0.3	+3.4±0.3
PEMO-MVN	-7.3±0.1	-4.2±0.3	+3.1±0.3

measured by Wagener et al. (1999c) was -7.1 ± 0.8 dB SNR. Hence, the stationary noise condtion was well predicted by simulations with all features. Using GBFB features resulted in the lowest simulation results (-8.2 ± 0.1 dB SNR). For the modulated noise, the picture changes considerably. Simulated SRTs ranged from -19 to 0 dB SNR depending on the employed feature set, where the empirical values measured by Wagener and Brand (2005) were on average -21.6 ± 2.0 dB SNR. The lowest simulation results and hence, those closest to the empirical data, were obtained with GBFB and SGBFB features, with -16.2 ± 0.5 dB SNR and -19.0 ± 0.4 dB SNR, respectively, followed by MFCC features with -15.0 ± 0.7 dB SNR. At the far end of the range, the use of LogMS and PEMO features resulted in simulated SRTs higher than in the respective stationary condition with -0.5 ± 0.3 dB and -3.7 ± 0.3 dB SNR, respectively.

The effect of modulation, which was defined as the difference in dB between the modulated (IRCA5) and the stationary (Olnoise) noise condition, was found to be -14.5 ± 2.2 dB for the empirical data. This means that for listeners with normal hearing it was much easier to recognize speech in the modulated noise condition than in stationary noise condition. Comparing the modulation effect with the LogMS feature set ($+6.3 \pm 0.3$ dB), which performed no modulation processing, the SGBFB-R-T feature set ($+6.2 \pm 0.3$ dB), which only performed the temporal modulation filtering, the SGBFB-R-S feature set (-6.7 ± 0.5 dB), which only performed the spectral modulation filtering, and the SGBFB-RR feature set (-10.8 ± 0.4 dB), which performed both, shows that spectral modulation processing alone accounts for the major part of the modulation effect and that temporal filtering alone has no effect. Deviations of simulated thresholds from the empirical data are further analyzed in Sec. 7.3.5.

7.3.4 Effect of back-end parameter variations

The simulation results with varied back-end parameters are depicted in Fig. 7.9. For the simultaneous masking experiment the average simulated thresholds and for the German Matrix sentence test the simulated SRTs are plotted depending on the varied back-end parameters for MFCC and PEMO features.

Generally, the smallest parameter value from the range of considered values which resulted in the lowest thresholds ± 1 dB was chosen if no reason existed not to do so. While for the Matrix sentence test, the words were best modeled with HMMs with 6 emitting states, for the simultaneous masking experiment it was sufficient to use HMMs with a single emitting state. The special states (START and STOP) were chosen according to the simulation results from the German Matrix sentence test because for the simultaneous masking experiment long special states (> 6 states) effectively narrowed the region to search for the target tone and hence improved the thresholds in an unwanted manner. Hence, the border effects were modeled best with HMMs with 6 emitting states. Reducing the amount of training data resulted in higher simulated thresholds, while increasing the amount of training data did not result in improvements of simulated thresholds exceeding 1 dB. It should be noted that the number of training samples per model guaranteed that each mean and each variance in the GMM was estimated from at least 96 samples, which

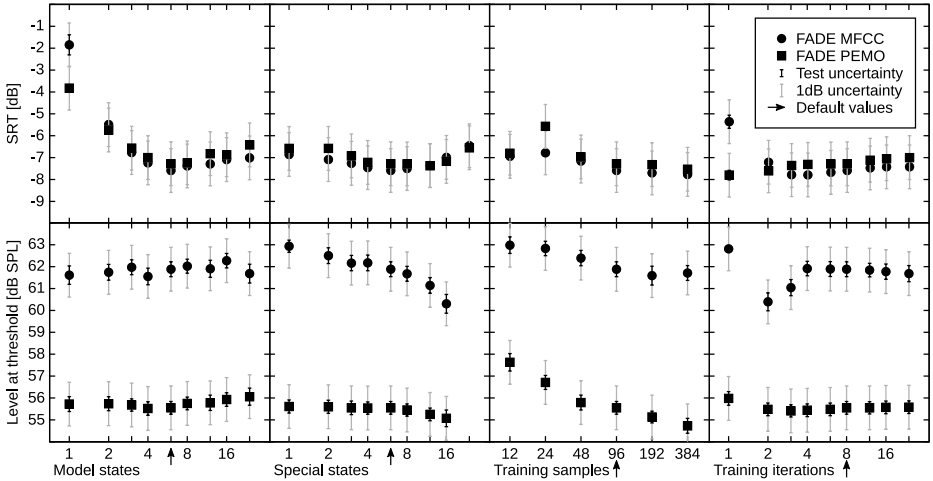

Figure 7.9: *Predicted SRTs (upper row) and tone detection thresholds in noise (lower row) from the back-end parameter variation experiment in dB SPL and dB SNR. The number of states per model, the number of states of the special models (START and STOP), the number of training samples per model, and the number of training iterations were varied over wide ranges of possible values. The predicted thresholds for the Matrix test in the test-specific noise condition and the average predicted thresholds for the tone-in-noise detection thresholds are plotted depending on the altered parameter values for both considered front-ends. The circles and squares indicate the results when using the MFCC and PEMO front-end, respectively. The arrows indicate the default parameter values. The small (partly hidden behind the markers) black error bars indicate the uncertainty due to finite number of testing samples, and the larger, gray error bars indicate the target precision of* ± 1 *dB.*

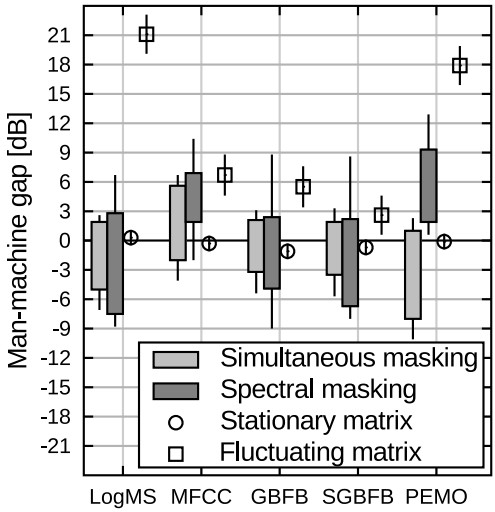

Figure 7.10: *Differences between simulated thresholds and empirical data depending on the feature set and experiment group. The difference is interpreted as the gap between human performance and machine performance; the lower the values, the smaller the gap, where positive and negative values indicate sub-human and super-human recognition performance, respectively. For the masking experiments, each of which has several conditions, the maximum and the minimum difference to the empirical data is depicted. The error bars indicate the uncertainty of the corresponding (minimum/maximum) value.*

was only the case if the corresponding HMM state occupied only one frame, i.e., the shortest possible duration of an HMM state. The number of training iterations was sufficient, with a security margin of factor 2, for all models to converge during the training procedure.

7.3.5 Man-machine gap

To get a comprehensive overview of the model fidelity depending on the employed feature set and experiment, the maximum and minimum differences from the empirical data is reported in Figure 7.10. While negative values indicate an over-estimation of the empirical thresholds,

positive values indicate an under-estimation. It should be noted that for human listeners no significant difference was found if the German Matrix test was presented in a closed-set or open response format (Warzybok et al., 2015b). The over all maximum can be interpreted as the remaining (unexplained) gap between human performance and machine performance. In this regard, the German Matrix sentence test in the modulated noise condition was the decisive condition, or very near ($< 1\,\mathrm{dB}$) to the decisive condition, for all feature types. Over all considered experiments, only the feature sets without spectral modulation processing, or in more general terms *across-frequency processing*, (LogMS, PEMO, SGBFB-R-T) resulted in under-estimated thresholds that were off by more than an order of magnitude ($> 10\,\mathrm{dB}$). The ASR features (MFCC, GBFB, and SGBFB) provided simulation results which came closer to the human performance or even exceeded human performance in some tasks. The simulation results which least under-estimated the empirical thresholds were obtained using SGBFB features, by under-estimating the empirical performance by no more than $2.6 \pm 2.0\,\mathrm{dB}$, followed by GBFB features with $5.5 \pm 2.1\,\mathrm{dB}$, and MFCC features with $6.9 \pm 3.5\,\mathrm{dB}$.

7.3.6 Effect of feature vector normalization

Considering the results in Tables 7.1, 7.2, and 7.3, the MVN was found to have a minor effect on the simulation results except for the LogMS features in the spectral masking experiment simulation, where the deviation of a simulated threshold was exceptionally high. The use of MVN did neither qualitatively improve the overall simulation fidelity with LogMS-MVN or PEMO-MVN features nor did its omission when using MFCC-MVN features.

7.4 DISCUSSION

It was shown that FADE enables the simulation of discrimination experiments of highly variable complexity using different feature vectors. The simplest experiment was a tone-in-noise detection task and the most complex the recognition of German Matrix sentences in a modulated noise condition. The feature vectors included traditional and robust ASR features as well as the output of a non-linear auditory model. The simulated thresholds were interpreted as predictions for the outcome of

the corresponding experiment when performed by listeners with normal hearing.

7.4.1 Interpretation of simulated thresholds

All simulated thresholds are reference-free (i.e., neither the deviation from a reference-signal-based "optimal detector" nor an empirical reference threshold was employed) and were obtained with a recognition system that was primarily constrained by the input signals and the signal representation. In comparison to many models of psychoacoustic performance, the approach to construct or train an "optimal detector" with prior knowledge about the exact temporal stimulus alignment (such as, e.g., employed by Dau et al. (1996a, 1997)) is replaced by a training phase of the Hidden Markov Model and the selection of the respective training condition yielding the lowest predicted threshold. This selection requires *feedback* about the recognition performance and is the only information with that FADE in its current version is provided and human listeners usually not. However, human listeners could probably guess the SNR at which they are listening. Hence, to better simulate the human recognition task, it seems worthwhile to investigate the possibility of taking the decision blindly in future work. It should be noted that the criterion for the decision on the optimal training data set is *recognition performance* and independent from any empirical data, as opposed to determining a fixed, e.g., training-test SNR offset, based on empirical data. The FADE approach, which decodes feature sequences instead of matching patterns, also models the uncertainty about the temporal alignment of the stimuli. Hence, it might be considered as more appropriate model of the human recognition process than an "optimal detector", which requires a-priori information that human listeners do not have access to. In comparison to state-of-the-art methods of robust ASR, the simulation of the German Matrix sentence test actually is an ASR experiment, but with most of the generic demands on a robust ASR system moved aside. That is to say, the ASR setup is not constructed to accommodate, e.g., generalization over speakers, noise conditions, reverberation, dialects, and other factors. Over a common ASR experiment, the approach to drastically reduce the number of those very broad demands has the advantage that it clearly shows when a feature set is not able to cope with a situation, like in the fluctuating noise condition of the German Matrix sentence test.

As a simulation with FADE is a very controlled A(S)R experiment, the same interpretation as in ASR is valid: the lower the threshold the "better" the system. In this context, thresholds below the corresponding empirical thresholds mean super-human recognition performance and thresholds above the corresponding empirical thresholds mean there is a gap in performance between the man and the machine, also referred to as the *man-machine-gap*. It should be noted that this interpretation is only possible because the thresholds with FADE are reference-free, objective thresholds and that this property translates naturally to the domain of psychoacoustic experiments.

While in the domain of ASR it is difficult to achieve (and hence predict) super-human performance because of its extensive demands, which result in a high variability of the signals to be recognized, in the domain of psychoacoustic experiments it is relatively easy to predict super-human performance because the trained detector stage (the HMMs in our case) can be highly specialized to the well-defined stimuli, which show less variability. This hypothesis is supported by the data in Figure 7.10, where for the speech recognition tasks no significant super-human performance was predicted, while for the tone detection tasks, some simulated thresholds were below the corresponding empirical thresholds. For current "optimal detector"-based psychoacoustic models, the additional a-priori information about the temporal alignment theoretically further facilitates achieving super-human performance predictions.

Even though the main prediction result of the current work concerns the threshold estimation discussed so far, more details of the FADE simulations might be considered to further validate the modeling of speech recognition and psychoacoustic tasks performed so far. For example, the slope of the psychometric function at the threshold could be derived from the recognition result map (RRM) and compared to empirical data. Likewise, the RRM could be evaluated for, e.g., each word group separately and word confusion matrices could be derived. Also, the selected training conditions could reveal differences between different feature sets.

7.4.2 Signal processing dependence of simulated thresholds

The simulated thresholds were found to depend on the employed feature set, where, in the speech recognition cases, the least variability was observed for the German Matrix sentence test in the test-specific noise

condition and the most variability was observed for the German Matrix sentence test in the modulated noise condition. In the latter, the least fitting thresholds (−0.5 dB SNR) were obtained with LogMS features while the best predicting thresholds (−19.0 dB SNR) were obtained with SGBFB features, spanning a range of almost two orders of magnitude (20 dB). In the tone-in-noise detection experiments the dependence on the feature set was not as pronounced as in the modulated noise condition of the German Matrix sentence test. As, apart from the feature set, nothing in the setup was changed, this finding confirms the hypothesis that the signal processing employed in the feature extraction process plays an important role in modeling auditory experiments.

Interestingly, the simulated thresholds for the German Matrix sentence test in the test-specific noise condition were not found to depend on the very different feature sets, i.e., PEMO and MFCC features, while the simulated thresholds in the modulated noise condition exposed the decisive shortcomings of some of the considered feature sets (cf. Sec. 7.4.5). Hence, the modulated noise condition was found to be the "critical" experiment to distinguish across the feature sets employed here.

Schädler and Kollmeier (2015a) observed in a robust ASR experiment that an ASR system using GBFB features outperformed one using MFCC features, and one using SGBFB features outperformed one using GBFB features. Further, one can assume that the LogMS features will generally not outperform MFCC features in robust ASR tasks as well. The same pattern was observed in the simulated thresholds of the modulated noise condition. Obviously, the most complex experiment of the current study, the German Matrix sentence test in the modulated noise condition, poses very similar basic demands on the employed feature set as in realistic robust ASR tasks. In future work, it could be investigated if this correspondence holds for different features and robust ASR tasks.

7.4.3 Required assumptions for ADE simulations

In comparison to current psychoacoustical modeling approaches, FADE poses comparatively few assumptions about the tasks and stimuli, i.e., the following assumptions must be valid in order to simulate an experiment with FADE.

Psychometric function The primary assumption is, that the goal of the experiment is to determine a point on a psychometric function.

The psychometric function needs to indicate the recognition rate on an auditory discrimination task depending on an independent variable which controls the difficulty of the task. The number of classes which have to be discriminated must be limited. In the current study the classes were either target and reference, or 50 different words of which 10 needed to be discriminated at a time, i.e., 1-out-of-2 and 1-out-of-10 discrimination tasks.

The same stimuli as in the original experiment As the basic idea is to estimate the lowest obtainable threshold given a certain task, a set of stimuli, and a signal representation (features), the signals used to perform the simulation must be the same that were used in the original experiment. More technically, the method to generate signals of different classes (e.g., target and reference) for different values of the independent variable must be provided. The signal representations must exhibit a certain variability which may be due to the signal itself (such as, e.g., external noise or other sources of variations within the provided signals) or due to a stochastic process in the feature extraction (such as, e.g., internal noise or uncertainty about the signal and which feature is best suited). For the experiments in the current work, the noise and speech signals caused sufficient variations, and the feature extraction was deterministic. The shortest stimulus used in the current study was a tone which lasted 5 ms, the longest was a word (the German word "achtzehn") which lasted about 900 ms. Technically, no hard limitations with respect to the stimulus length exist.

Observable effects due to signal processing The observable effect must originate from the interaction between the stimuli and the signal processing involved in the feature extraction, where the stimuli incorporate the task requirements and the signal processing the limitations of the human auditory system. This condition expresses the requirement that, differences in the stimulus which are not apparent in the signal representation cannot be detected by the recognition system and will hence not result in different thresholds.

7.4.4 Generalization of the FADE approach

One set of parameters was shown to suffice for a variety of experiments and features (cf. Fig. 7.9). The criterion to determine these parameters was the lowest obtainable thresholds and hence, they were independent of the empirical data of the considered tasks. These parameters also worked well in the simulation of the experiments which are not included in Fig. 7.9, i.e., the spectral masking experiment and the German Matrix sentence test in the modulated noise condition. Hence, the FADE approach generalized well over the considered experiments and features. The fact that a single set of parameters was sufficient for a variety of complex tasks and different types of features provides evidence that the underlying approach might be appropriate to simulate more experiments and that other features can be incorporated as well to model an even larger variety of experiments with the same set of parameters.

7.4.5 Across-frequency processing and relation to temporal processing

The data from Table 7.3 indicates that a correct direction of the modulation effect, (i.e., a reduction in SRT by about 14.5 dB in humans due to modulations imposed on the noise) was only found for feature sets which incorporated some kind of across-frequency processing. For example, when extracting MFCCs, the DCT was calculated in the spectral dimension of the LogMS and hence MFCCs integrated over the whole spectral bandwidth. With GBFB and SGBFB features the LogMS was spectrally band-pass filtered. With these feature sets improved thresholds were found in the modulated noise condition. However, an opposite effect, i.e., the predicted thresholds increased in the modulated noise condition compared to the stationary noise condition, was observed for LogMS and PEMO features, of which the spectral bands are assumed to be independent. The SGBFB-RR features, a reduced set of SGBFB features, allowed to perform either only the temporal modulation processing (SGBFB-R-T) or only the spectral modulation processing (SGBFB-R-S). The simulated thresholds with these tailored feature sets showed that the temporal processing alone (SGBFB-R-T) did not show an appropriate modulation effect, while the spectral processing alone (SGBFB-R-S) was sufficient to obtain an improved threshold in the modulated noise condition. With

the set up implemented in this study it was not possible to explain the modulation effect without some kind of across-frequency processing.

A representation which allows the reliable detection of local spectral maxima based on, e.g., slope or curvature, instead of absolute values, which have to be relied on if no across-frequency comparison is performed, could probably help the back-end in decision-taking. Hence, it seems possible that at least some kind of across-frequency processing, in its most explicit form the spectral modulation processing performed by the SGFB-R-S feature set, is required to recognize speech in fluctuating noise. If true, this finding might have far-reaching consequences for any system (biological or technical) with the intention to recognize human speech, as it puts the common understanding that speech can be processed in independent frequency bands into question. For example, it might be desirable to preserve spectral modulation patterns rather than temporal modulation patterns in signal processing strategies of hearing devices if preserving speech intelligibility in non-stationary background noise is a declared intention.

Another yet unresolved question is if the spectral and temporal modulation processing in the human auditory system interact with each other or if they are separate processes. Schädler and Kollmeier (2015a) observed that no spectro-temporal interaction in the modulation filtering, i.e., inseparable spectro-temporal filters, was needed to outperform MFCC and GBFB features in an ASR system employed in acoustically adverse conditions which included spectrally, temporally and spectro-temporally modulated noise. This observation is supported by the thresholds of the modulated noise condition that were simulated in this study. In Figure 7.10, the simulated thresholds obtained with SGBFB features were among the most suitable for explaining the empirical data. This could indicate that the SGBFB features might be a reasonable model of the auditory processing in the human auditory system and, if so, hint that spectral and temporal modulations in the human auditory system might be processed separately.

7.5 CONCLUSIONS

The most important findings of this work can be summarized as follows:

— A simulation framework for auditory discrimination experiments (FADE) was successfully employed to simulate, and hence, predict the outcome of a broad range of auditory detection experiments with an increasing complexity while requiring fewer assumptions compared to traditional modeling approaches.

— A single set of general parameters was determined which was used to simulate all experiments from the basic tone-in-noise detection experiment to the complex speech-in-modulated-noise recognition task.

— Across-frequency processing was found to be crucial to predict the improved speech reception threshold in modulated noise conditions over stationary noise conditions.

— Of all considered signal representations, the Gabor filter bank based features with some across-frequency processing, most notably GBFB and SGBFB features, provide the most suitable model of human performance across the considered experiments.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft SFB TRR 31 “The active auditory system” and the Cluster of Excellence Grant “Hearing4all”.

8 General Conclusions

The main goal of this thesis was to improve the robustness of automatic speech recognition systems by integrating auditory signal processing principles.

In Chapter 1, a general approach to this goal was outlined. In Chapters 2, 3, and 4, it was successfully tackled by integrating auditory-motivated, physiologically-inspired spectro-temporal modulation filters which extract so called Gabor filter bank features (GBFB) into the front-end of a standard ASR system. Further, in Chapter 5, the GBFB features were reduced in complexity by decomposing the spectro-temporal modulation filtering into separate spectral and temporal processes, introducing the separable Gabor filter bank (SGBFB) features. In Chapter 6, following the initially outlined approach (cf. Chapter 1), a recognition experiment was established in which human listeners and a standard ASR system perform equally well: The simulation of a matrix sentence test in stationary noise. This approach was then successfully extended to simulating tone-in-noise detection experiments and a matrix sentence test in a fluctuating noise condition in Chapter 7.

Spectro-temporal modulation features for robust ASR

The results in Chapters 2, 3, and 4 showed that the direct encoding of spectro-temporal modulation patterns in the feature vectors can improve the robustness of an ASR system, even in medium-size vocabulary (≈ 5000 words) tasks. The results in Chapter 5 confirmed this finding and showed that the feature extraction can be decomposed into two separate filtering processes, a spectral and a temporal one. This indicates that no interaction between the spectral and the temporal signal is required.

Auditory spectro-temporal modulation processing

The results in Chapter 7 confirmed that spectro-temporal features are required to achieve human recognition performance even in a very simple and controlled ASR task – at least from an ASR point-of-view – if a

fluctuating background noise was used. From the considered feature extraction algorithms, SGBFB features were the only ones with which a standard recognition system could keep up with human recognition performance in all of a variety of recognition tasks. This result provides evidence that SGBFB features are *auditory* features and could be interpreted as a faint hint that spectral and temporal processing in the human auditory cortex might be separate processes.

Simulation of auditory discrimination experiments

In Chapter 7, it was shown that the proposed simulation framework for auditory discrimination experiments (FADE) is already able to simulate basic tone-in-noise detection as well as simple speech recognition tasks. FADE constitutes the foundation for a series of increasingly more complex auditory discrimination experiments to detect, isolate and tackle potential shortcomings of ASR systems with high precision. For example, across-frequency processing was found to be crucial to recognize speech in a modulated noise condition. The proposed framework can also be seen as a new, performance-oriented approach to modeling the human perception of sound in auditory discrimination experiments, which offers important advantages over traditional approaches. The simulated results are *reference-free,* i.e. no empirical data is required to calibrate them, which enables independent, model-based predictions. The proposed approach is more realistic in that *neither frozen nor clean signals* but the same signals which are available to the human listeners, are used for modeling, i.e. the model has no signal-based advantage because the task is as difficult as for the human listeners. Further, the approach encourages to make *as few assumptions as possible,* because every unmatched assumption adds to the risk that the system will eventually fail on or be unable to simulate one of the increasingly complex recognition tasks. The approach is *universal* in the sense that the same model parameters can be used to simulate a broad range of auditory discrimination experiments. The possibility to evaluate a feature set across tasks which put diverse demands on it helps to prevent or easily identify over-fitting to a specific task.

In contrast to current psychoacoustic models, exact pattern matching ("matched filter") is avoided in favor of a statistically trained recognition model (HMM) on the back-end side. The most important difference between the traditional approach of pattern matching and the proposed

approach which uses sequence decoding is that the former assumes perfect prior knowledge about the temporal alignment of the stimuli while for the latter, like for human listeners, determining the temporal alignment is part of the task. The inherent accuracy limit of current psychoacoustic models is not determined by the uncertainty about the time-dependent succession of the assumed "hidden states" (as assumed by the HMM in the FADE approach) but rather by a limitation/degradation of the signal representation (e.g. "internal noise" or smearing of the "internal image"). Note, however, that this assumed degradation requires an adjustment to empirical data in a reference condition, which is prone to over-fitting. The additional prior information about the temporal alignment provides current psychoacoustic models with a performance reserve which might enable sub-optimal models of auditory signal processing to outperform human listeners and could possibly discourage further investigation on important signal processing matters. The performance-oriented perspective taken by the FADE approach allows to compare models of signal processing and human listeners on more similar terms, i.e. based on the same a-priori information. The man-machine gap on a group of tasks is then easily determined by taking the maximum deviation from the empirical thresholds, as described in Section 7.3.5. By taking the maximum, it is impossible to average out a flaw, and attention is always drawn to the condition where the largest gap between human and machine performance is visible. This interpretation can guide future research endeavors by clearly demanding solutions to the most imminent problem, i.e. the largest man-machine gap, first.

Future work

> *Future robust ASR systems, models of speech intelligibility, and psychoacoustic models should converge.*

A common paradigm for hearing and speech research based on machine learning would serve all disciplines involved; with more faithful models, more accurate predictions, more robust features, and eventually a solid common knowledge base. To allow convergence, this new paradigm requires:

— implemented (instead of descriptive) models of auditory principles to enable the simulation of experiments

— limited information on the signal level, i.e., using the same signals as in empirical studies

— human-like performance on all auditory discrimination tasks in a direct comparison to empirical data

For well-designed experiments it enforces a fair comparison of man and machine performance, disentangling the *auditory task* from the *auditory model*. Further, it penalizes over-specific assumptions, e.g. overfitting of models to specific tasks, and rewards universally valid approaches/assumptions, e.g. using spectro-temporal representations. The framework for simulating auditory discrimination experiments (FADE) is aimed to implement this paradigm and could serve as a common simulation/prediction platform for future work. It bridges the gap from psychoacoustic models over speech intelligibility models to robust ASR and allows the interchange of signal representations (features), which account for the largest part of implementable knowledge about the human auditory system. A lively exchange of thoughts and approaches in hearing and speech research would accelerate the development of better hypotheses about the human auditory system and speech perception. FADE, although not perfect because it automatically selects the best training condition based on the simulation results, constitutes a major step towards a universal model for hearing and speech research in that it shares as many components/parameters as possible. As a next step, the basis of auditory discrimination tasks and the basis of auditory signal representations (features) will need to be extended to get an as-broad-as-possible view on the strengths and shortcomings of existing models. Time will tell which auditory signal processing strategies turn out indispensable, and which untenable. By then, the different research disciplines might have agreed on a common, universal auditory signal representation.

Bibliography

ANSI. American National Standard: "Methods for the Calculation of the Articulation Index", 1970.

ANSI. American National Standard: "Methods for Calculation of the Speech Intelligibility Index", 1997.

Auditec. CD101RW2, 2006. URL http://www.auditec.com. Audio CD.

Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633, 2013.

Barker, J., Marxer, R., Vincent, E., and Shinji, W. The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU) 2015*. IEEE, 2015.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007.

Bouvrie, J., Ezzat, T., and Poggio, T. Localized spectro-temporal cepstral analysis of speech. In *Proceedings of ICASSP 2008*, pages 4733–4736. IEEE, 2008.

Bradley, J. S., Sato, H., and Picard, M. On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America*, 113(6):3233–3244, 2003.

Brand, T. and Kollmeier, B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, 111(6): 2801–2810, 2002.

Carnegie Mellon University. The CMU pronouncing dictionary, 2007. URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Castro Martinez, A. M., Moritz, N., and Meyer, B. T. Should deep neural nets have ears? The role of auditory features in deep learning approaches. In *Proceedings of Interspeech 2014*, pages 2435–2439, 2014.

Chi, T., Ru, P., and Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

Cole, R. A., Noel, M., Lander, T., and Durham, T. New telephone speech corpora at CSLU. In *Proceedings of Eurospeech 1995*. ISCA, 1995.

Cooke, M. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573, 2006.

Cooke, M. and Scharenborg, O. The Interspeech 2008 consonant challenge. In *Proceedings of Interspeech 2008*, pages 1765–1768. ISCA, 2008.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20 (1):30–42, 2012.

Dau, T., Püschel, D., and Kohlrausch, A. A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *The Journal of the Acoustical Society of America*, 99(6):3615–3622, 1996a.

Dau, T., Püschel, D., and Kohlrausch, A. A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements. *The Journal of the Acoustical Society of America*, 99 (6):3623–3631, 1996b.

Dau, T., Kollmeier, B., and Kohlrausch, A. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, 102(5): 2892–2905, 1997.

Davis, S. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4): 357–366, 1980.

De La Torre, A., Peinado, A. M., Segura, J. C., Pérez-Córdoba, J. L., Benítez, M. C., and Rubio, A. J. Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3):355–366, 2005.

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

Derleth, R. P. and Dau, T. On the role of envelope fluctuation processing in spectral masking. *The Journal of the Acoustical Society of America*, 108(1):285–296, 2000.

Domont, X., Heckmann, M., Joublin, F., and Goerick, C. Hierarchical spectro-temporal features for robust speech recognition. In *Proceedings of ICASSP 2008*, pages 4417–4420. IEEE, 2008.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. ICRA Noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *International Journal of Audiology*, 40(3):148–157, 2001.

Ellis, D. P. W. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. URL http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/.

ETSI. Standard: "201 108 v1.1.3" Speech processing transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, 2003.

ETSI. Standard: "202 050 v1.1.5" Speech processing transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, 2007.

Ewert, S. D. and Dau, T. Characterizing frequency selectivity for envelope fluctuations. *The Journal of the Acoustical Society of America*, 108(3): 1181–1196, 2000.

Ezzat, T., Bouvrie, J., and Poggio, T. AM-FM demodulation of spectrograms using localized 2D max-Gabor analysis. In *Proceedings of ICASSP 2007*, volume 4, pages 1061–1064. IEEE, 2007a.

Ezzat, T., Bouvrie, J. V., and Poggio, T. Spectro-temporal analysis of speech using 2-D Gabor filters. In *Proceedings of Interspeech 2007*, pages 506–509. ISCA, 2007b.

Fletcher, H. and Galt, R. H. The perception of speech and its relation to telephony. *The Journal of the Acoustical Society of America*, 22(2): 89–151, 1950.

Fraunhofer IDMT, 2014. URL http://www.idmt.fraunhofer.de/en/Service_Offerings/products_and_technologies/q_t/sip-toolbox.html.

Gillick, L. and Cox, S. J. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of ICASSP 1989*, pages 532–535. IEEE, 1989.

Gramss, T. Word recognition with the feature finding neural network (FFNN). In *Proceedings of the Workshop on Neural Networks for Signal Processing*, pages 289–298. IEEE, 1991.

Green, D. M. and Swets, J. A. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

Hagerman, B. Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*, 11(2):79–87, 1982.

Heckmann, M., Domont, X., Joublin, F., and Goerick, C. A closer look on hierarchical spectro-temporal features (HIST). In *Proceedings of Interspeech 2008*, pages 894–897. ISCA, 2008.

Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

Hermansky, H. and Fousek, P. Multi-resolution RASTA filtering for TANDEM-based ASR. In *Proceedings of Interspeech 2005*, page 361–364. ISCA, 2005.

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

Hermansky, H. and Sharma, S. Temporal patterns (TRAPS) in ASR of noisy speech. In *Proceedings of ICASSP 1999*, volume 1, pages 289–292. IEEE, 1999.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. Rasta-plp speech analysis technique. In *Proceedings of ICASSP 1992*, volume 1, pages 121–124. IEEE, 1992.

Hermansky, H., Ellis, D. P. W., and Sharma, S. Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of ICASSP 2000*, volume 3, pages 1635–1638. IEEE, 2000.

Hochmuth, S., Brand, T., Zokoll, M. A., Castro, F. Z., Wardenga, N., and Kollmeier, B. A spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology*, 51 (7):536–544, 2012.

Hochmuth, S., Kollmeier, B., Brand, T., and Jürgens, T. Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests. *International Journal of Audiology*, 54 (sup2):62–70, 2015.

Hohmann, V. and Kollmeier, B. The effect of multichannel dynamic compression on speech intelligibility. *The Journal of the Acoustical Society of America*, 97(2):1191–1195, 1995.

Holube, I. and Kollmeier, B. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *The Journal of the Acoustical Society of America*, 100(3): 1703–1716, 1996.

ISO. Standard: "226: 2003 acoustics – normal equal-loudness-level standard", 2003.

Jepsen, M. L., Ewert, S. D., and Dau, T. A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, 124(1):422–438, 2008.

Jørgensen, S. and Dau, T. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130(3):1475–1487, 2011.

Jørgensen, S., Ewert, S. D., and Dau, T. A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):436–446, 2013.

Jürgens, T. and Brand, T. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The Journal of the Acoustical Society of America*, 126(5):2635–2648, 2009.

Kanedera, N., Arai, T., Hermansky, H., and Pavel, M. On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, 28(1):43–55, 1999.

Kleinschmidt, M. Methods for capturing spectro-temporal modulations in automatic speech recognition. *Acta Acustica united with Acustica*, 88(3):416–422, 2002b.

Kleinschmidt, M. Localized spectro-temporal features for automatic speech recognition. In *Proceedings of Eurospeech 2003*, pages 2573–2576. ISCA, 2003.

Kleinschmidt, M. and Gelbart, D. Improving word accuracy with gabor feature extraction. In *Proceedings of Interspeech 2002*, pages 25–28. ISCA, 2002a.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, 54(sup2):3–16, 2015.

Kryter, K. D. Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697, 1962.

Leijon, A. Estimation of sensory information transmission using a hidden Markov model of speech stimuli. *Acta Acustica United with Acustica*, 88(3):423–432, 2002.

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

Lochner, J. and Burger, J. The influence of reflections on auditorium acoustics. *Journal of Sound and Vibration*, 1(4):426–454, 1964.

Ludvigsen, C. *The use of objective measures to predict the intelligibility of hearing aid processed speech. Recent developments in hearing instrument technology*, pages 81–94. Scanticon, Kolding, Denmark, 1993.

Mesgarani, N., Slaney, M., and Shamma, S. A. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3): 920–930, 2006.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, 123(2):899–909, 2008.

Mesgarani, N., Thomas, S., and Hermansky, H. A multistream multiresolution framework for phoneme recognition. In *Proceedings of Interspeech 2010*, pages 318–321. ISCA, 2010.

Meyer, B. T. and Kollmeier, B. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Communication*, 53(5):753–767, 2011a.

Meyer, B. T., Brand, T., and Kollmeier, B. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. *The Journal of the Acoustical Society of America*, 129(1):388–403, 2011b.

Meyer, R. M. and Brand, T. Comparison of different short-term speech intelligibility index procedures in fluctuating noise for listeners with normal and impaired hearing. *Acta Acustica united with Acustica*, 99 (3):442–456, 2013.

Moore, B. C. J., Alcántara, J. I., and Dau, T. Masking patterns for sinusoidal and narrow-band noise maskers. *The Journal of the Acoustical Society of America*, 104(2):1023–1038, 1998.

Moritz, N., Anemuller, J., and Kollmeier, B. Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments. In *Proceedings of ICASSP 2011*, pages 5492–5495. IEEE, 2011.

Moritz, N., Schädler, M. R., Adiloglu, K., Meyer, B. T., Jürgens, T., Gerkmann, T., Kollmeier, B., Doclo, S., and Goetze, S. Noise robust distant

automatic speech recognition utilizing NMF based source separation and auditory feature extraction. In *Proceeding of CHiME Workshop 2013*, pages 1–6, Vancouver, British Columbia, Canada, 2013.

Nadeu, C., Macho, D., and Hernando, J. Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication*, 34(1):93–114, 2001.

Nemala, S. K. and Elhilali, M. Relevant spectro-temporal modulations for robust speech and nonspeech classification. *The Journal of the Acoustical Society of America*, 127(3):1817–1817, 2010.

Ozimek, E., Warzybok, A., and Kutzner, D. Polish sentence matrix test for speech intelligibility measurement in noise. *International Journal of Audiology*, 49(6):444–454, 2010.

Patterson, R. D. and Moore, B. C. J. Auditory filters and excitation patterns as representations of frequency resolution. *Frequency Selectivity in Hearing*, pages 123–177, 1986.

Pearce, D. and Hirsch, H.-G. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of Interspeech 2000*, pages 29–32. ISCA, 2000.

Qiu, A., Schreiner, C. E., and Escabí, M. A. Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition. *Journal of Neurophysiology*, 90(1):456–476, 2003.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise. *The Journal of the Acoustical Society of America*, 126(6):3236–3245, 2009.

Schröder, J., Moritz, N., Schädler, M. R., Cauchi, B., Adiloglu, K., Anemüller, J., Doclo, S., Kollmeier, B., and Goetze, S., S. On the use of spectro-temporal features for the IEEE AASP challenge 'Detection and classification of acoustic scenes and events'. In *Proceeding of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2013*, pages 1–4. IEEE, 2013.

Schwartz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L., and Zavaliagkos, G. Comparative experiments on large vocabulary speech recognition. In *Proceedings of the Workshop on Human Language Technology*, pages 75–80. Association for Computational Linguistics, 1993.

Schädler, M. R. and Kollmeier, B. Normalization of spectro-temporal Gabor filter bank features for improved robust automatic speech recognition systems. In *Proceedings of Interspeech 2012*, pages 1812–1815. ISCA, 2012b.

Schädler, M. R. and Kollmeier, B. Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition. *The Journal of the Acoustical Society of America*, 137(4):2047–2059, 2015a.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *The Journal of the Acoustical Society of America*, 131(5):4134–4151, 2012a.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology*, 54(sup2):100–107, 2015b.

Stadler, S., Leijon, A., and Hagerman, B. An information theoretic approach to predict speech intelligibility for listeners with normal and impaired hearing. In *In Proceedings of Interspeech 2007*, pages 1345–1348. ISCA, 2007.

Steeneken, H. J. and Houtgast, T. A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326, 1980.

Stevens, S. S. On the psychophysical law. *Psychological review*, 64(3): 153, 1957.

Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. J. The importance for speech intelligibility of random fluctuations in "steady" background noise. *The Journal of the Acoustical Society of America*, 130(5):2874–2881, 2011.

Stone, M. A., Füllgrabe, C., and Moore, B. C. J. Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, 132(1):317–326, 2012.

Søndergaard, P. L. and Majdak, P. The auditory modeling toolbox. In *The technology of binaural listening*, pages 33–56. Springer, 2013.

Tchorz, J. and Kollmeier, B. A model of auditory perception as front end for automatic speech recognition. *The Journal of the Acoustical Society of America*, 106(4):2040–2050, 1999.

Varga, A. and Steeneken, H. J. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.

Verhey, J. L., Dau, T., and Kollmeier, B. Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model. *The Journal of the Acoustical Society of America*, 106(5):2733–2745, 1999.

Vertanen, K. Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments. Technical report, Cavendish Laboratory, University of Cambridge, Cambridge, UK, 2006.

Viemeister, N. F. Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America*, 66(5):1364–1380, 1979.

Viikki, O. and Laurila, K. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147, 1998.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In *Proceedings of ICASSP 2013*, pages 126–130. IEEE, 2013a.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU) 2013*, pages 162–167. IEEE, 2013b.

Wagener, K., Brand, T., and Kollmeier, B. Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil I: Design des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test – Part I: Design of the Oldenburg sentence test). *Zeitschrift für Audiologie*, 38:4–15, 1999a.

Wagener, K., Brand, T., and Kollmeier, B. Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil II: Optimierung des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test – Part II: Optimization of the Oldenburg sentence tests). *Zeitschrift für Audiologie*, 38:86–95, 1999b.

Wagener, K., Brand, T., and Kollmeier, B. Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil III: Evaluation des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test – Part III: Evaluation of the Oldenburg sentence test). *Zeitschrift für Audiologie*, 38:44–56, 1999c.

Wagener, K. C. and Brand, T. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. *International Journal of Audiology*, 44(3):144–156, 2005.

Warzybok, A., Rennies, J., Brand, T., Doclo, S., and Kollmeier, B. Effects of spatial and temporal integration of a single early reflection on speech intelligibility. *The Journal of the Acoustical Society of America*, 133 (1):269–282, 2013.

Warzybok, A., Zokoll, M., Wardenga, N., Ozimek, E., Boboshko, M., and Kollmeier, B. Development of the russian matrix sentence test. *International Journal of Audiology*, 54(sup2):35–43, 2015a.

Warzybok, A., Brand, T., Wagener, K. C., and Kollmeier, B. How much does language proficiency by non-native listeners influence speech audiometric tests in noise? *International Journal of Audiology*, 54 (sup2):88–99, 2015b.

Wesker, T., Meyer, B. T., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In *Proceedings of Interspeech 2005*, pages 1273–1276. ISCA, 2005.

Wong, L. L. N., Ho, A. H. S., Chua, E. W. W., and Soli, S. D. Development of the Cantonese speech intelligibility index. *The Journal of the Acoustical Society of America*, 121(4):2350–2361, 2007.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., et al. *The HTK book (for HTK version 3.1)*. 2001.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., et al. *The HTK book (for HTK version 3.4)*. 2006.

Zhao, S. Y. and Morgan, N. Multi-stream spectro-temporal features for robust speech recognition. In *Proceedings of Interspeech 2008*, pages 898–901. ISCA, 2008.

Zhao, S. Y., Ravuri, S. V., and Morgan, N. Multi-stream to many-stream: Using spectro-temporal features for ASR. In *Proceedings of Interspeech 2009*, pages 2951–2954. ISCA, 2009.

Zwicker, E. *Frequency Analysis and Periodicity Detection in Hearing*, chapter Masking and psychological excitation as consequences of the ear's frequency analysis, pages 376–396. Sijthoff Leiden, 1970.

Danksagung

Ich möchte mich ganz herzlich bei allen bedanken, die mich durch mein Leben begleiten: Bei Isa und Nico, bei meinen Familien, bei meinen Freunden und Kollegen und bei Birger, der mir diese Arbeit ermöglicht hat.

Desweiteren möchte ich mich bei jedem Leser meiner Arbeit für das Interesse bedanken.

Vielen Dank :)