From Acoustic Mismatch Towards Blind Acoustic Model Selection in Automatic Speech Recognition

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Thomas Winkler
aus
Göttingen

Bonn 2012

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr.-Ing. Christian Bauckhage
2. Gutachter:	Prof. Dr. Michael Clausen
Tag der Promotion:	17.04.2013
Erscheinungsjahr:	2013

Zusammenfassung

Akustische Störungen und akustische Fehlanpassungen sind zwei der kritischsten Einflüsse auf die Automatische Spracherkennung. Ein Sprachsignal, das in einer bestimmten akustischen Umgebung aufgenommen wurde, kann deutlich unterschiedliche Signalcharakteristiken zu einem Sprachsignal der selben Äußerung aus einer anderen akustischen Umgebung aufweisen. Akustische Merkmale, die für eine Automatische Spracherkennung verwendet werden, sind nicht ideal und beinhalten zusätzlich zu den entscheidenden Merkmalen für eine Spracherkennung auch solche Störeinflüsse. Während diese Störeinflüsse an sich bereits eine Verschlechterung der Spracherkennung in gestörten akustischen Umgebungen bewirken, ist eine akustische Fehlanpassung der Spracherkennung ein noch gravierenderes Problem. Eine solche Fehlanpassung tritt dann auf, wenn ein System auf Sprache in bestimmten akustische Bedingungen trainiert wurde, aber unter anderen akustischen Bedingungen eingesetzt wird.

In dieser Arbeit analysieren wir detailliert Einflüsse verschiedenartiger Störquellen und damit verbundener Fehlanpassungen von Hintergrundgeräuschen, über Mikrofoncharakteristiken bis hin zu Kodier- und Kanalübertragungseinflüssen. Dabei werden die unterschiedlichen Einflüsse in Hinblick auf Veränderungen des Sprachsignals, der extrahierten Merkmale und der Spracherkennungsergebnisse sowohl unter angepassten als auch fehlangepassten Bedingungen genau analysiert und bewertet. Für diesen Zweck stellen wir verschiedene Evaluationskorpora vor. Zwei der Korpora wurden für die genannten Untersuchungen entwickelt. Dabei ist insbesondere der MoveOn-Korpus zu nennen, der unterschiedlich gestörte und verzerrte Sprachsignale zur Verfügung stellt und in dieser Arbeit eine zentrale Rolle spielt. Design und Entwicklung dieses Korpus werden daher detailliert vorgestellt.

Bei der Untersuchung der genannten akustischer Störungen zeigt sich unter anderem, dass bereits geringe Störungen unter Umständen signifikante Auswirkungen auf die extrahierten Merkmale und die Automatische Spracherkennung haben können. Zudem sind die Auswirkungen der Störungen auf Signale und insbesondere Merkmale sehr unterschiedlich und stark abhängig von verschiedenen Parametern. Aufgrund der Komplexität bestimmter Einflüsse ist ein allgemeiner Ansatz zur Kompensation der Störungen kaum möglich.

Da jedoch sowohl die Merkmale als auch die akustischen Modelle, die für die Spracherkennung verwendet werden, unweigerlich einen Teil der Informationen über die akustische Störung mit aufnehmen, stellen wir einen neuen Ansatz für eine multi-modell-basierte Spracherkennung vor. Dieser Ansatz bestimmt automatisch, welches akustische Modell aus einer Auswahl von verschiedenen gut angepassten Modellen am besten für das aktuelle Sprachsignal geeignet ist. Dabei werden für die Auswahl nur die Merkmale und die akustischen Modelle ohne zusätzliches Wissen und ohne nähere Annahmen über die Störeinflüsse berücksichtigt, weshalb wir diesen Ansatz als blinde akustische Modellauswahl (“blind acoustic model selection”) bezeichnen. Zur Geschwindigkeitsoptimierung unseres Ansatzes präsentieren und evaluieren wir ein Verfahren zur Reduktion der akustischen Modelle. In der Praxis werden mit unserem Verfahren der Modellauswahl bei geeigneten akustischen Modellen etwa gleich gute bis leicht verbesserte Ergebnisse gegenüber vergleichbaren multi-konditionalen Modellen erzielt.

Abstract

Acoustic distortion and acoustic mismatch are two of the most critical aspects influencing automatic speech recognition. A speech signal recorded from a speaker in a certain acoustic environment compared to a signal from the same utterance recorded under different acoustic conditions can have very different characteristics. Acoustic features used for automatic speech recognition are not ideal and also incorporate such acoustic influences in addition to the information relevant for speech recognition. While distortion of the signal caused by difficult acoustic conditions already reduces the recognition accuracy, additional acoustic mismatch in case of a system trained in one particular acoustic condition and used under different acoustic conditions further decreases the performance.

In this work we offer a detailed analysis of the influences of various sources of acoustic distortion and acoustic mismatch from additive noise, microphone characteristics towards coding and transmission channel effects. We evaluate and understand their influence on the speech signal, the extracted features, and the speech recognition performance in matched and mismatched conditions. For this purpose we introduce several speech and noise corpora appropriate for evaluating these aspects. Two of these corpora are purposely designed and recorded for the presented evaluations. In particular the MoveOn Corpus offers an evaluation corpus of realistic noisy speech generally useful for research on robust automatic speech recognition beyond the scope of this thesis. Thus, design decisions and corpus development are detailed for this corpus.

Based on the presented speech corpora we analyse various acoustic conditions and show the effects of even small changes in the speech signal, which can have a significant influence on the extracted speech features and the recognition performance. The changes in features can be quite manifold and are dependent on the particular distortion and other parameters as we will discuss in detail. Thus, such changes are usually difficult to compensate by a universal approach.

As features and acoustic models commonly used for automatic speech recognition inevitably inherit part of the information on the acoustic conditions we propose and evaluate a new multi-model approach selecting a best matching set out of several sets of well adapted acoustic models solely based on the extracted features and the acoustic models. We call this approach blind acoustic model selection as it works completely blind neither incorporating additional knowledge nor any particular assumption about the type of acoustic distortion. For improved processing speed we further suggest to use a compact representation of each set of acoustic models instead of the full set. The results indicate that the theoretical performance clearly outperforms commonly used multi-conditional acoustic models. In case of an appropriate selection of the sets of acoustic models comparable or even improved results compared to multi-conditional acoustic models are also achieved in practice with our proposed approach.

Acknowledgements

I would like to thank all my colleagues who supported me in many ways. Thank you for providing the necessary space to prepare and finalise my thesis as well as for feedback and new insights in our many discussions. My sincere thanks are given in particular to my family, my partner, and my friends who believed in me and fully supported me during the long time of my thesis with many long and exhausting days. I know that you would have liked to spend more time with me during this time, and I would have loved to do so, too. Only because of your patience and understanding these and all the other lines in this thesis exist. Thank you very much!

Ich möchte mich sehr herzlich bei all meinen Kollegen bedanken, die mich in vielerlei Hinsicht unterstützt haben. Danke für die nötigen Freiräume, die ihr mir gegeben habt, um meine Arbeiten umzusetzen und abzuschließen, und für die Anmerkungen und neuen Einsichten aus unseren zahlreichen Diskussionen. Mein herzlicher Dank geht insbesondere auch an meine Familie, meine Freundin und meine Freunde, die an mich geglaubt haben und mich während der Zeit meiner Doktorarbeit mit den oft langen und anstrengenden Tagen uneingeschränkt unterstützt haben. Ich weiß, dass ihr in dieser Zeit oft gerne mehr Zeit mit mir verbracht hättet, und ich hätte es sehr gerne auch getan. Nur durch Eure Geduld und Euer Verständnis konnten diese und all die anderen Zeilen in dieser Arbeit entstehen. Vielen herzlichen Dank!

Contents

Chapter 1

Introduction

For more than 50 years automatic speech recognition (ASR) and its application challenges scientists and engineers. Many advances in pattern recognition and related fields led to first commercial applications about two decades ago. More data, new technologies and approaches, as well as improving processing power of computer systems enabled many new applications of ASR making it more and more popular nowadays. While in the 1990s mainly call centre applications with ASR for simple dialogue interaction on telephone speech were used, more complex systems for large vocabulary dictation tasks like Nuance's *Dragon Naturally Speaking* came up at the very end of the 20th century. New technologies in the area of ASR and an increasing amount of data available for training improved such existing systems but also enabled new applications from Google's voice search to Apple's personal assistant *Siri* in the second half of the young new millennium. Still, even in controlled environments and for clean speech, for example, in case of dictating a text into a high quality microphone in an office room, the existing systems do not work perfectly. In even more difficult environments on a busy street using a mobile phone, the performance often even drops significantly and the acceptance of speech driven applications suffers enormously. Thus, the effort in researching and developing robust algorithms for ASR, which are less affected by harsh acoustic environments, has increased in the last two decades.

Background noise and other sources of distortion degrading the quality of the speech signal can be considered as the major reason for the significant differences of the speech recognition performance in very noisy acoustic environments compared to a quite environment. While humans cope rather well with low to medium degradations of human speech, machines suffer much more from such influences. A central issue in machine-based speech recognition as a classification problem is the extraction of appropriate features, which shall describe all characteristics relevant for the classification of the spoken content but neglect all others. In case of typical speech features used for ASR today, we can observe that in particular the suppression of irrelevant information for the classification process, which includes information about the acoustic environment or the speaker, is far from being ideal. Considering that exactly the same type of features used for ASR are often also used for a general acoustic classification or in speaker recognition point to the fact that indeed much information about these influences is present in these features. As such features work well for clean conditions and well trained systems and as no significantly better features have been found until now, they are still widely used today and research often focuses on approaches that try to deal with the insufficiency of the features by removing or compensating for parts of the irrelevant information. Unfortunately, the influence of various distortions on common speech features is often complex showing manifold and often non-linear effects highly dependent on the type of distortion. Thus, no reliable and universal approach has been found so far. While distortion generally degrades the signal's quality, and thus, the recognition performance, an additional mismatch between training data and speech data encountered during recognition caused by different characteristics of distortion is even much more critical and must be avoided whenever possible. A proper adaptation to a certain acoustic condition is often the only satisfactory solution yielding significant improvements of the performance of the speech recognition system in most situations.

Most research in the area of robust speech recognition focuses on a certain aspect of acoustic distortion making use of certain assumptions to reduce the effect of some source of distortion. This often leads to approaches that work considerably well in similar situations with similar distortion, but sometimes even decrease the performance in different acoustic situations where the assumptions are not true any more. Work considering and comparing the different effects caused by various distortions is scarce, as researchers often focus on one or two aspects only. Very often also simulated distorted speech data instead of realistic data is used for evaluations. Such data is usually much easier to collect in a sufficient amount and further offers controlled and known characteristics of distortion as they are influenced during simulation. But various effects are known that indicate that simulated distorted speech is not necessarily very similar to realistic distorted speech. While these effects are known, hardly any work evaluates the differences between both types of data, and experiments are often performed with simulated sets only.

We want to pick up and examine several of the problems described above in the following thesis. We will present the design and development of a realistic speech and noise database enabling the evaluation on realistic speech data as well as a comparison of realistic and simulated noisy speech. Based on the developed database we provide an exhaustive analysis of various sources of acoustic distortion in the context of ASR. Our analysis covers major influences from background noise, microphone channels, towards transmission channel effects including hardware and coding/decoding effects. Common approaches for noise reduction and mismatch compensation are evaluated on the different distortions and compared to a well adapted speech recognition system. We further compare a close to realistic simulation of distorted speech with the realistic reference data to learn about similarities and differences in their characteristics and about weak points in current simulation approaches. The knowledge gained by these evaluations motivates the last part of the thesis introducing a new approach of ASR using multiple acoustic models for a successful speech recognition in various acoustic conditions. This approach is solely based on the acoustic information in the extracted speech features and the acoustic models and shows promising results comparable or even superior to a common approach of multi-conditional acoustic models in many situations.

1.1 Scientific Goals

In this thesis we will analyse and identify challenges for ASR due to acoustic mismatch in the speech features caused by acoustic distortion. In an extensive evaluation of various types of acoustic distortion we will point out the influences on common speech features and the ASR performance for each of the major sources separately. We will further identify weak points of certain assumptions and simplifications commonly assumed for noisy speech simulation and noise reduction in the context of robust ASR. This requires a speech and noise database, which was not available before, thus introducing the sub-goal of creating such a database suitable for the evaluations on acoustic mismatch and robust ASR in this work. The results of the previous investigations point to the final goal: we want to provide a new approach for ASR in multiple and realistic acoustic conditions considering the provided knowledge of the previous evaluations. This approach must be easily extendible to arbitrary new conditions and must work in various realistic conditions without using any particular assumption or knowledge about these conditions during recognition.

We can detail these main goals providing more detailed tasks and sub-goals that we plan to accomplish in this thesis:

- For our evaluations in this work we will provide a purposely designed and developed speech and noise corpus. Speech and noise must be realistic, simulations during corpus design and develop-

ment are prohibited. This corpus must enable a separate analysis of various sources of distortion and must provide realistic noise samples to simulate and compare simulated noisy speech and realistic noisy speech in the context of robust ASR.

- Our following evaluation of major influences of different sources of distortion on a speech signal, speech features and ASR includes mainly three major sub-goals:
  - First, we separately analyse and understand the influences of three major sources of distortion on the signal and the ASR process. The three sources include background noise, microphone channel characteristics, and transmission channel characteristics (including hardware and coding/decoding effects).
  - We evaluate and understand the limitations of common approaches for noise reduction and mismatch compensation. We further carve out the difference in performance between common mismatch approaches and a well adapted system for ASR.
  - We identify specific challenges and provide recommendations for robust ASR based on the previous results. We incorporate the knowledge of the changes in speech features caused by the different distortions and the severe influences on the ASR process caused by mismatch introduced that way.
- We further analyse and point out chances and limitations of simulation approaches of certain of the evaluated acoustic conditions and distortions. While such approaches enable controlled experiments on one or another source of distortion, simplifications and wrong assumptions can cause significant differences compared to realistic distorted signals.
- With the knowledge and recommendations derived from the previous chapter, we want to achieve a new multi-model approach for multi-conditional ASR. This approach should fulfil the following sub-goals:
  - The approach must be able to cope with various acoustic conditions. The acoustic conditions can be very different and we do not make any restrictions about the acoustic domain. The conditions should be known by the system. That means we generally have sufficient data or a prepared set of acoustic models for each of the acoustic domains the system should work in.
  - The approach must be able to easily incorporate new acoustic conditions, provided that, again, sufficient data or a prepared set of acoustic models for the new conditions is available to the system. All processes from adding new conditions to speech recognition should be unsupervised.
  - The approach should not rely on any particular assumption about the type of distortion. Preferably, no additional information beyond the speech features and the acoustic models is considered.
  - The performance in terms of speed and accuracy of the recognition process should not be influenced significantly compared to the case of one domain adapted set of acoustic models.
  - Even though we focus on known acoustic conditions avoiding acoustic mismatch whenever possible, mismatch reduction approaches will also be considered for cases of unavoidable mismatch.

To achieve the aforementioned scientific goals in this work, we will evaluate relevant aspects from acoustic mismatch towards blind acoustic model selection in automatic speech recognition.

1.2 Related Publications

Parts of this work build upon previously published material by the author of this thesis. These publications will briefly be introduced here.

In [1, 2] we described the process of designing, collecting and preparing a speech and noise corpus (the MoveOn Corpus) that offers two very different, synchronously recorded microphone channels: one throat microphone channel synchronously recorded with two close-talk microphones. The throat microphone is worn around the neck picking up vibrations from the larynx, while the two close-talk microphones are installed left and right from the mouth of the speaker in a motorcycle helmet. This setup of microphones was used to record speech and noise on the motorcycle in the real environment on the road as well as in a noise free office environment.

The MoveOn Corpus was used for several experiments evaluating the general performance of an ASR system in these different environments, for domain-specific acoustic models as well as for acoustic mismatch with and without mismatch compensation. The evaluation published in [3] showed the effect of the different channels and distortions on the ASR performance. Furthermore, we could show that mismatch compensation is capable of improving the recognition results compared to recognition without compensation for mismatched conditions. We could also show that mismatch compensation is far from being perfect and is still just a fallback providing rather poor results when compared to a well adapted system.

In [4] we further evaluated the effect of the different channel characteristics of the two types of microphones in the MoveOn Corpus. Compared to a close-talk microphone a throat microphone has a rather poor spectral coverage of the signal lacking certain frequencies especially in the higher frequency bands. On the other hand, the signal of the throat microphone is almost noise free (concerning additive background noise), as it directly picks up the vibrations of the larynx as opposed to airborne sound. Both effects lead to different performances in ASR which were presented in detail in this publication.

Another relevant distortion in robust ASR is additive noise. Often speech with artificially added noise is used as it is easier to create and to influence the degree of distortion, but certain aspects are rather different to realistic additive noise. One major aspect is the Lombard effect that influences the way we speak in noisy conditions. This effect changes certain speech characteristics with the intention of improving the comprehensibility of the spoken content. This effect is difficult to simulate as it is highly variable from speaker to speaker and is also influenced by the type of background noise, so that usually a simplified noise model assuming pure additive noise is used to create simulated data. In [5] we compared realistic noisy speech with simulated noisy speech based on the characteristics of the realistic reference data. We could show that the influence of the background noise was in any case present and critical for the ASR performance, but that simulated and realistic data were not that similar at all. Thus, realistic data is crucial to develop a system for noise robust speech recognition in a universal context.

Finally, we also created a test corpus based on clean speech with a step by step simulation of the TETRA (Terrestrial Trunked Radio) channel including close to realistic TETRA channel data sent through actual TETRA hardware. This corpus is used to evaluate certain channel effects beyond the microphone channel including influences of TETRA hardware, coding/decoding algorithms and simple low-pass filtering. In [6, 7] we published a brief description of the creation of this corpus as well as parts of the presented evaluations.

The summarised work motivated our concept of blind acoustic model selection. Here, we introduce a new concept for a multi-model approach with acoustic model selection solely based on the extracted speech features and sets of acoustic models for ASR. In addition to this approach we show the effect of using a compact representation instead of the full sets of acoustic models for acoustic model selection. Such a compact representation significantly increases the speed of acoustic model selection while hardly

affecting the ASR performance in terms of accuracy. The concept of blind acoustic model selection is detailed in [8] and is further elaborated in this thesis.

1.3 Structure of Thesis

The main part of this thesis is divided into four parts. The first part (Chapter 2) is an introductory chapter describing common approaches and challenges for ASR. We briefly outline the history of ASR and describe a common statistical approach for ASR based on HMMs and MFCCs. A major challenge for ASR is acoustic distortion and variability. We picture the main sources of distortion and variabilities in this chapter giving an idea of the complexity of the possible distortions. Afterwards we present numerous existing approaches for robust ASR, which aim at reducing the effects of one or several of the presented distortions on the ASR performance. Some of these approaches are also applied or adapted in the subsequent chapters. Finally, we introduce common measures for the performance of ASR systems that we will also use in our evaluations.

In Chapter 3 we present the different speech corpora that we will use in the following chapters for evaluations on acoustic distortion and robust ASR. The AURORA Project Database 2.0 is briefly summarised first. We use this corpus as a reference corpus for our evaluations as it is widely used in the scientific community. We further created two additional corpora purposely designed and developed for evaluations in this thesis. The first is the TETRA (Terrestrial Trunked Radio) Corpus enabling evaluations on the step by step degradation of the TETRA radio transmission. This corpus offers several simulated steps of a TETRA radio transmission of the speech signal as well as a realistic transmission of the signal via TETRA hardware. Afterwards we detail the MoveOn Motorcycle Speech Corpus from design to implementation also including some statistics and baseline results for ASR. The design of the MoveOn corpus allows for extensive evaluations planned in this work. Major characteristics, strengths and weaknesses of each corpus concerning the following evaluations are summarised at the end.

We use the MoveOn and TETRA Corpus in Chapter 4 to analyse and evaluate various aspects of speech distortion. An introductory section about an integrated ASR system shows the general influence and importance of the main parts of an ASR system on the overall performance. We further elaborate these aspects in the following sections analysing and evaluating the effects of background noise and microphone channel on the MoveOn Corpus first. The synchronously recorded signals from two different microphone techniques further enable a comparison of the influences of the microphone channel in the context of ASR. A final evaluation on the TETRA Corpus analyses the effects of the transmission channel for realistic and simulated distorted data in more detail by applying a step by step degradation of the signal. The results from these evaluations provide valuable information on the variable characteristics of the different distortions and enable us to identify problems and recommendations for robust ASR.

This knowledge about distortion and mismatch motivates Chapter 5, where we introduce a new concept for acoustic model selection from multiple acoustic models. This approach is evaluated and compared to a common alternative approach of multi-conditional acoustic models. Furthermore, we introduce and evaluate approaches for feature normalisation for mismatch reduction based on our approach of acoustic model selection. In our evaluations of the presented approaches and in the final conclusion we will show and discuss the potentials of our concept and its advantages and disadvantages compared to multi-conditional training.

In a final conclusion in Chapter 6 we briefly recapitulate the evaluated aspects and stress results from this thesis, which are also relevant in future research in robust ASR. The scientific achievements of this thesis are summarised in Chapter 7.

Chapter 2

Automatic Speech Recognition (ASR)

Automatic speech recognition (ASR) for controlled tasks and "clean" (noise free) environments is often considered to be solved. Still, recognition performance of such systems is only close to perfection and significantly drops for more adverse environments or for tasks which are quite different from the one the system was built for. Plenty of approaches were developed in the last decades to make ASR less prone to any kind of distortion and acoustic mismatch, which are two of the main reasons for a decreasing performance. ASR incorporating any approach to improve the recognition performance even in adverse environments is called *robust* ASR. Most of the robust ASR systems are built up on typical, non-robust systems extending or modifying one or several aspects of the speech recognition process.

This chapter will build the foundation of the following work presented in this thesis. As we want to identify and measure the influences of acoustic distortion on ASR with a state-of-the-art system, we need to understand the general concept of such a system. We will see the importance of appropriate training and evaluation data, which will be relevant in Chapter 3, for investigating in main aspects of acoustic distortion. Furthermore, knowledge about typical sources of acoustic distortion as well as their influence on the speech signal, the speech features and the recognition process is crucial. Often, such interaction between speech and distortion is simplified and one or several aspects are just neglected when trying to simulate or deal with acoustic distortion in robust ASR. This background is helpful to understand the following chapters and is especially relevant for Chapter 4 where we will isolate specific influences of different sources of acoustic distortion on ASR and analyse the complex of problems simulating such distortion. All the previous discussion points to the central issue of robustness in ASR. Plenty of approaches were developed in the last decades that try to compensate for mismatch caused by one or several sources of distortion but are not successful enough in most realistic scenarios. Thus, robustness of speech recognition is one of the most relevant directions of research in ASR since the task of recognising clean speech under controlled acoustic conditions is practically solved. Some of the main directions of research on this topic of robust ASR are described here and relevant approaches for each direction are presented. This provides a foundation for Chapter 4, where we will also discuss the shortcomings of some approaches, and Chapter 5, where we propose a new multi-model approach motivated by the chapter before and the experiences from the various approaches presented here.

In the following we present a brief history of ASR showing the development, increased complexity, and new challenges in its field of research. Afterwards a state-of-the-art statistical approach for ASR based on hidden Markov models (HMMs) and mel-frequency cepstral coefficients (MFCCs) commonly used today is summarised. Further, we will discuss one of the major issues in ASR today, the problem of distortion and acoustic mismatch caused by different influences, before we will present an overview of prominent and new approaches for robust ASR dealing with one, several or all of the introduced types of distortion. A common measure for describing the performance of a speech recognition process and commonly used for evaluating ASR systems is introduced at the end of this chapter.

2.1 History and Challenges of ASR

The very beginning of automatic speech recognition started with the recognition of isolated words, usually by template-based methodologies. In the 1950s the first approaches mainly considered a vocabulary of up to 10 words (usually isolated digits or monosyllabic words, [9, 10]) spoken by a single speaker. This number increased to a vocabulary in the order of 10–100 words in the 1960s, when time-normalisation techniques (e.g. [11]) and dynamic programming methods including dynamic time warping (DTW; [12]) were introduced to ASR, so that variations due to the speaking rate could be compensated. Still, for such small vocabulary tasks, matching was done by simple acoustic-phonetic properties.

With advanced normalisation techniques and new algorithms of dynamic programming (in particular the Viterbi algorithm - [13, 14]), the introduction of linear predictive coefficient (LPC; [15]) analysis and advances in pattern recognition applied to speech recognition, more flexible systems for connected digits and continuous speech became feasible in the 1970s. Such systems were also able to tackle medium vocabulary tasks with about 100–1000 words (e.g. CMU's Harpy; [16]).

Hidden Markov models (HMMs; [17, 18]) as a key technology and the introduction of stochastic language models ([19]) in the 1980s, brought speech recognition to the next step enabling large vocabulary continuous speech recognition systems with a vocabulary beyond 1000 words. Furthermore, mel-frequency cepstral coefficients (MFCCs; [20, 21]) with its first and second order derivatives ([22]), which were already proposed in the 70s for speaker recognition, became more and more popular for ASR. These technologies from the 80s with certain improvements from the last 30 years of research are still used today in most state-of-the-art speech recognisers.

With the chance of solving large vocabulary tasks of continuous speech, the step from speech recognition to speech understanding became more and more relevant and first methods for stochastic language understanding were developed in the 1990s. State of the art stochastic model based systems could further be improved by an increasing amount of available speech and language data for training improved stochastic acoustic and language models. Also in the 1990s the research area of robust ASR got boosted. While large vocabulary continuous speech recognition was possible for certain, noise free environments at that time, speech recognition performance dropped significantly for background noise, channel mismatch and other influences in the audio signal. First major advances in this field were amongst others maximum likelihood linear regression (MLLR; [23]) or parallel model combination (PMC; [24]). Simple applications of speech recognition in telephone networks — mainly for simple dialogue systems — started to become very popular in the 90s.

In the last 12 years since the beginning of the new millennium, large vocabulary continuous speech recognition for clean and planned speech like broadcast news further developed and reached results of up to 95% word accuracy. Major reasons for improved ASR performance is an increasing amount of available acoustic and textual data for training of the stochastic models as well as the development of algorithms capable of handling such large amount of data. Modern systems are trained on hundreds or thousands of hours of transcribed speech and billions of words of domain-specific texts. Thus, more and more systems beyond telephone dialogue systems come up that provide a sufficient quality to be commercialised (e.g., systems for spoken document retrieval [25], or dictation tasks [26]). In combination with new developments in other fields new applications like Google Voice Search [27] became possible, or with improvements in language understanding personal assistant systems on mobile phones (e.g. Apple's Siri[1]) now reach the customer.

Nevertheless, the development summarised above considers mainly clean and planned speech applic-

[1] http://www.apple.com/iphone/features/siri-faq.html

ations and often lacks sufficiently good results in many other fields of application with more difficult acoustic conditions. Thus, research in ASR nowadays has a strong focus on algorithms and methods to improve speech recognition accuracies in case of speech variations — in particular spontaneous speech — as well as background noise and other sources of signal distortion. The more severe these variations and distortions are, the higher the word error rates become limiting the possible applications. Thus, research in robust speech recognition with advanced adaptation techniques, speaker variability reduction, as well as channel and noise compensation interest more and more scientists and engineers. An early approach of the last decade tackling the issue of robust speech recognition was the ETSI robust front end for robust feature extraction standardised in 2002. Other relevant approaches for robust speech recognition in the last twelve years include amongst many others missing feature based approaches [28], various Wiener filtering methods [29], multi microphone approaches [30] and various model and feature adaptation methods [31, 32]. Still, the problem of speech recognition under difficult conditions is not solved yet. Spontaneous or distorted speech poses a problem even for advanced systems incorporating recent approaches for robust speech recognition with results often far from the accuracies for clean, planned speech.

While a complete overview would be out of scope, we refer to more detailed work on history and milestones of modern ASR in [33–35].

2.2 Statistical Approach to ASR

The goal of every ASR process is the transformation of spoken utterances into some kind of textual representation suited for one of the tasks described in the previous section. Several different approaches to achieve this goal exist today. Most concepts build up on a probabilistic approach to identify either consecutive sub word units (phoneme, syllable, etc.) or words. Such an approach uses features extracted from short time frames of a signal to calculate the acoustic probabilities for the sequence of frames matching a sequence of phonemes defined in the acoustic models. The acoustic likelihoods are usually re-weighted by lexical probabilities from a language model (likelihoods of words and word combinations) resulting in the final overall likelihoods of the possible word sequences.

In detail, the speech signal is segmented into several short frames, and the acoustic features for each frame are extracted. These features ideally provide relevant information for classifying different sounds of a language describing the spoken content and ignoring all information not relevant for this task. Typically, linear predictive coefficients (LPCs) or mel-frequency cepstral coefficients (MFCCs) are used in state-of-the-art systems. Based on a sequence of feature vectors, each acoustic entity that should be recognised can be described in an acoustic model. For modelling these entities, usually hidden Markov models (HMMs) are used that are able to cover dynamic characteristics in terms of varying lengths of the states of modelled entities, for example caused by different speaking styles. HMMs model processes by quasi stationary states and possible transitions (with certain transition probabilities) between these states. The states of each HMM are defined by probability distributions of all possible observations represented by feature vectors. A sufficient amount of manually transcribed sample utterances is required as training data to learn the probabilities for each HMM. Typical probability distributions are either single Gaussian distributions or Gaussian mixtures. Gaussian mixtures improve the coverage of acoustic varieties causing higher variations in the features, but require more samples per modelled entity, and thus, increase the demand for training data. During acoustic decoding the likelihoods for all possible states and transitions are calculated for the sequence of feature vectors of the unknown utterance. This results in a lattice of possible HMM sequences with the probability for each HMM.

In addition to the acoustic models, language models are usually incorporated in the recognition pro-

cess. A language model is trained from large amounts of textual data calculating probabilities of the occurrences of single words and probabilities of word combinations. Thus, we can include this knowledge during recognition re-calculating the probabilities of the lattice determined during acoustic decoding. This usually results in lattices of possible word sequences. Finally, the best hypothesis is estimated following the path in the lattice with the maximum likelihood. To achieve an optimal performance for a speech recogniser, both the language model and the set of acoustic models should preferably be trained on data from the targeted domain of the ASR system.

In the following (Section 2.2.1) we will detail the feature extraction step for MFCCs, as they are widely used in ASR. MFCCs provide very good results for controlled acoustic environments with no other types of features found so far, which significantly outperform MFCCs. We will further describe HMMs for acoustic modelling in speech recognition in Section 2.2.2, which can still be considered to be standard in typical ASR systems. A brief introduction to statistic language models is given in Section 2.2.3 complementing the acoustic models in state of the art systems. The acoustic and language probabilities provided by both types of models will be fused as described in the section about speech decoding (Section 2.2.4). While alternative approaches are also often considered in scientific work today, the presented statistic approach for ASR with MFCCs, HMMs, and stochastic language models is still one of the main approaches used in state-of-the-art ASR. Further information about stochastic based speech recognition can be found in [35–37].

2.2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

Various types of acoustic features exist, which can usually be used in HMM-based speech recognition as well as in completely different setups like neural networks. In this work we will focus on MFCCs as they are still one of the most commonly used types of features for ASR, especially when using a HMM-based approach for acoustic modelling. MFCCs can be considered as the amplitudes of the spectrum of the logarithmic mel-scaled short time spectrum of a signal. The general steps of MFCC extraction are as follows:

1. Short time Fourier transform of the signal
2. Mapping of the energies of the frequencies to the mel scale
3. Calculation of the logarithm of the mel-scaled energies of frequencies
4. Discrete cosine transform of the mel-scaled energies of frequencies
5. Taking the first N or all resulting amplitudes as MFCC features

Usually, additional short time energy and the derivatives of these static coefficients are added to the final feature vector.

This work flow of a standard MFCC feature extraction is also shown in Figure 2.1. First, a discrete speech signal is divided into overlapping short time segments or frames (usually with 25 ms length and a step size of about 10 ms). When we apply the commonly used Fast Fourier Transform to calculate the discrete frequencies, even for periodic signals so called leakage to other frequencies occurs due to usually non-periodic components introduced by cutting off the signal at the beginning and end of the frame. Thus, window functions are usually used that weight the samples closer to the edges of the frame close to or exactly to zero. The selection of the window function finally is a compromise between minimising leakage, on the one hand, and distortions of the frequencies caused by weighting of the samples, on the other hand. The value $x_n^{(m)}$ of a windowed signal for the frame at position m can

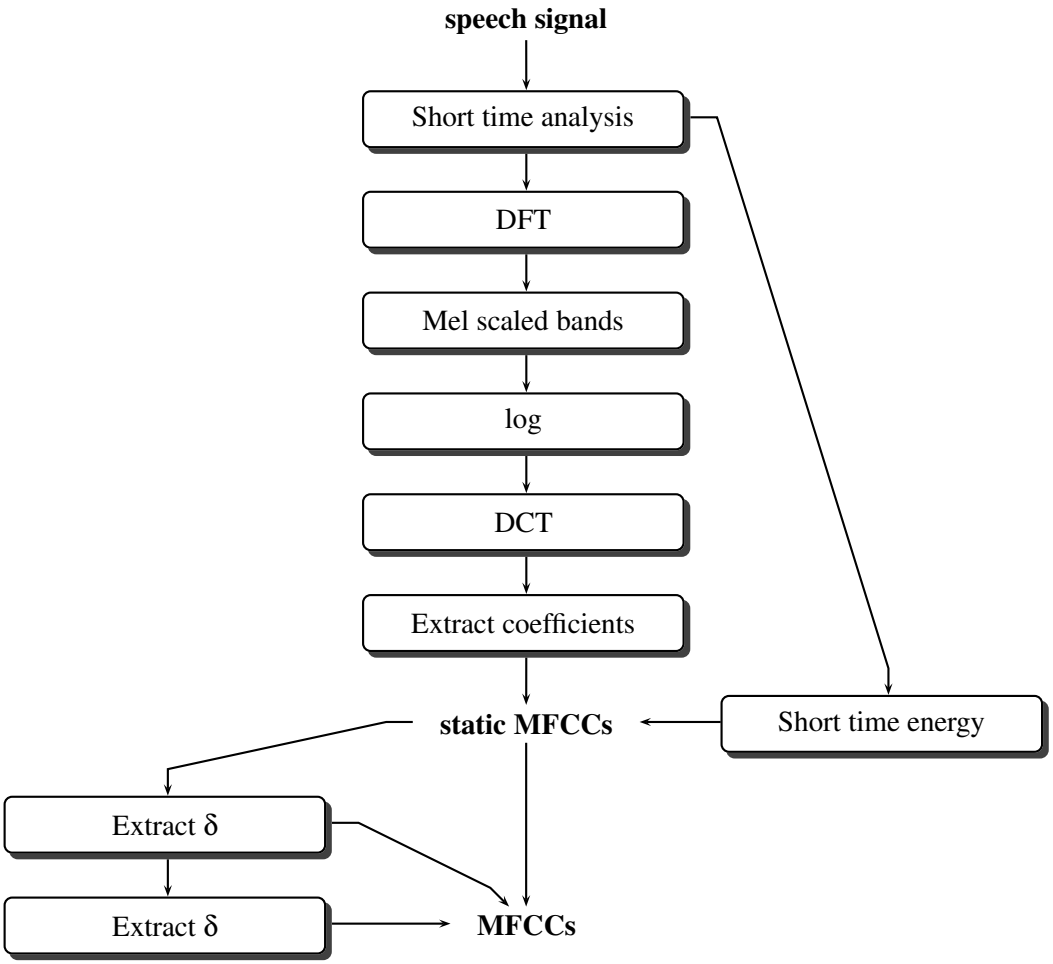


Figure 2.1: Work flow of MFCC feature extraction. A short time discrete Fourier transform (DFT) calculates the time frequency representation of a speech signal. The energies of the frequency bins are mapped to mel scale. The discrete cosine transform (DCT) of the logarithmic mel-scaled energies provides the mel-frequency cepstral coefficients. The first 12 coefficients plus short time energy and first and second order derivatives are commonly used for ASR.

be calculated from the n^{th} sample x_n of the discrete signal by a multiplication $x_n^{(m)} = x_n w_{n-m}$ with the respective value of a window function w_n of a length of N samples. Most typical functions for MFCC extraction include Hann (Equation 2.1) and Hamming windows (Equation 2.2).

$$w_n^{(HN)} = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right), & \text{if } 0 \leq n \leq N-1 \\ 0, & \text{if } n < 0 \text{ or } n \geq N \end{cases} \tag{2.1}$$

$$w_n^{(HM)} = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & \text{if } 0 \leq n \leq N-1 \\ 0, & \text{if } n < 0 \text{ or } n \geq N \end{cases} \tag{2.2}$$

The windowed samples of each frame are then transformed into the frequency space by using the discrete Fourier transform (DFT). The transform of the windowed signal is also called time-discrete short time Fourier transform (STFT). The value for frequency bin ν of the frame at position m is defined as follows:

$$X_\nu^{(m)} = \sum_{n=-\infty}^{\infty} x_n w_{n-m} e^{-2\pi i \nu (n-m)/N} \tag{2.3}$$

Only the magnitudes $|X_\nu^{(m)}|$ of the frequency bins are used, as in case of short window durations of typically 25 ms in ASR the information provided by the phase is assumed to be little.[2]

A mel-scaled filter bank is applied to the magnitudes of the frequency representation of each frame to approximate the logarithmic frequency response of the human auditory system [39]. The mel scale is designed to copy human perception of pitches judged to be equally distant from one another. A common formula to calculate the mel-frequency m from a frequency in Hertz f is as follows:

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.4}$$

The mel filter bank (Figure 2.2) provides higher resolution at low frequencies compared to higher frequencies similar to human perception.

The filter bank output of each bandpass filter provides the values to calculate the short time energy $e_k^{(m)}$ of the respective frequency group k. The weight $\eta_{k\nu}$ for frequency bin ν and filter bank k is derived from the triangular filters (compare Figure 2.2). The energies of the mel-scaled frequencies are calculated as follows:

$$e_k^{(m)} = \sum_{\nu=0}^{N-1} \eta_{k\nu} |F_\nu^{(m)}|^2 \tag{2.5}$$

Usually the logarithm of the energy is calculated to get a measure similar to loudness, which is much closer to human perception than the linear energy. The values of the logarithmic band energy are transformed by the discrete cosine transform (DCT) to receive the cepstrum of the mel log power spectrum. Introduced by Bogert et al. in [40] the "spectrum of the spectrum" that we get by applying the DCT on a spectrum, is called cepstrum. The discrete coefficients of the cepstrum are related to frequency-like bins called quefrencies. The q^{th} coefficient $c_q^{(m)}$ of the cepstrum of frame m is calculated from the K mel log spectral values $\log e_k^{(m)}$ by the DCT as follows:

[2] While some more recent work ([38]) indicate that there actually might be more useful information in the phase than anticipated, it is still common practice in ASR to discard the phase information.

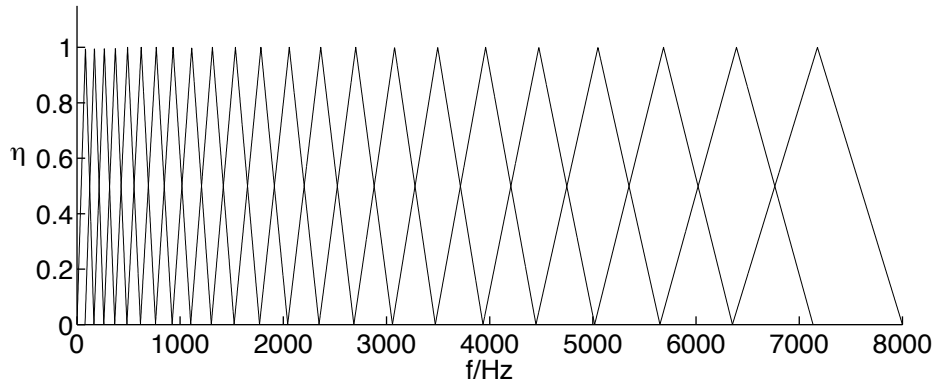


Figure 2.2: Mel filter bank with 23 bands and triangular filters. The presented example shows the mel-scaled centre frequencies for a frequency range from 0 to 8kHz. The weight $\eta_{k\nu}$ for filter k and frequency bin ν is 1 at the centre frequency of each filter bank and reaches 0 at the centre frequencies of the neighbouring filter banks.

$$c_q^{(m)} = \sum_{k=0}^{K-1} \log e_k^{(m)} \cos \frac{\pi q(2k+1)}{2K}, \quad q = 0,1,2,\ldots,K-1 \tag{2.6}$$

The DCT has a deconvolutional characteristic. Influences of the harmonics of the pitch, for example, which are approximately periodic in the mel-frequency scale and not relevant for ASR, will be present in the upper coefficients of the cepstrum, while formant information, which is relevant for ASR, will be found in the lower coefficients. Thus, only the amplitudes $c_q^{(m)}$ of the lower log quefrencies q of the discrete cepstrum are used as features. The zeroth coefficient $c_0^{(m)}$ is the sum of pitch related loudness (log energy) and is usually replaced by the short time energy directly calculated from each frame of the speech signal.

Temporal aspects of spoken language can be covered by MFCCs in several ways. The most common approach is using derivatives (mainly first and second order) of the static features. The calculation of the derivatives from the time discrete static features can be done in various ways. A simple approach is to use the difference of the static features τ frames before and after the current window m:

$$\Delta c_q^{(m)} = c_q^{(m+\tau)} - c_q^{(m-\tau)} \tag{2.7}$$

A usually better way to estimate the derivatives is regression as outliers at positions $m+\tau$ and $m-\tau$ have a lower negative impact on the results. For regression the features of all τ frames before and after frame m are considered:

$$\delta c_q^{(m)} = \frac{\sum_{j=-\tau}^{\tau} j c_q^{(m+j)}}{\sum_{j=-\tau}^{\tau} j^2} \tag{2.8}$$

The second order derivatives are calculated in the same way but based on the first order derivatives $\Delta c_q^{(m)}$ or $\delta c_q^{(m)}$ instead of the static features c_q^m.

A common MFCC feature vector contains the twelve static cepstral coefficients ($c_1^{(m)}$ to $c_{12}^{(m)}$). These coefficients contain the lower quefrency bins containing relevant information about formant structures and are usually sufficient. The zeroth coefficient $c_0^{(m)}$ is typically replaced by the short time energy of the frame, which is added to the feature vector instead. The first and second order derivatives of

the twelve cepstral coefficients and the short time energy are calculated and concatenated resulting in a 39 dimensional feature vector. Other combinations of static and dynamic MFCCs are possible and sometimes used.

2.2.2 Hidden Markov Models (HMMs)

Hidden Markov models (HMMs) for ASR were introduced and became a key technology of state-of-the-art ASR systems in the 1980s [17, 18]. HMMs enable modelling of time varying processes like speech and thus are widely used for ASR. A Markov chain named after Andrey Markov is a random process with a Markov property, meaning that the process is memoryless and the next state only depends on the current state. In detail, the Markov chain has a finite number N of discrete states $Q = \{s_1, \ldots, s_N\}$ and considers a discrete stochastic process $q = q_1, \ldots, q_T$ with $q_T \in Q$. The stochastic process has certain transition probabilities $P(q_t|q_{t-1})$ from q_{t-1} to q_t, which are not dependent on previous states as the process is memoryless. The transition probabilities can be joined in a matrix $A = [a_{ij}]_{N\times N}$ with $a_{ij} = P(q_t = s_j|q_{t-1} = s_i)$, with $a_{ij} \geq 0$ and $\sum_j a_{ij} = 1$. Furthermore, each state has a probability $\pi_i = P(q_1 = s_i)$ that it is the initial state in the process, with $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. The probabilities of the N states can be joined in an N-dimensional vector π.

Now a second process generates a certain output dependent on the current state of the stochastic process based on the finite set of possible outputs $V = \{v_1, \ldots, v_K\}$. The sequence of outputs can be observed, while the states in case of *hidden* Markov models cannot be observed directly. The generation of the observed sequence $O = o_1 \ldots o_T$ can be described by a second stochastic process with $P(o_t|q_t)$ only dependent on the current state q_t of the first process. So the observation probability of output v_k (seen in observation o_t) in case of finite state s_j is $b_{jk} = b_j(v_k) = P(o_t = v_k|q_t = s_j)$, with $b_{jk} \geq 0$ and $\sum_k b_{jk} = 1$. The observation probabilities can be stored in a matrix $B = [b_{jk}]_{N\times K}$. A hidden Markov process can then be defined by the number of states N, the set of outputs V of size K and the stochastic parameters $\lambda = (\pi, A, B)$. In case of speech recognition, the finite set of outputs V is replaced by continuous probability distributions.

While such a HMM describes a process and its generated visible observations, it can also be used to estimate the underlying generative but hidden states. Given a certain observation sequence, the initial state probability, the observation probabilities for each state and output as well as the transition probabilities of the HMM, the probably $P(O|\lambda)$ that an observation O is generated by a HMM λ, can be determined. This concept is used in ASR as detailed in the following sections.

HMMs in Speech Recognition

As opposed to many other applications where a transition to a variety of states is possible, HMMs in ASR allow forward transitions only — often even permitting transitions to any other state than the current and the next state. This is also shown in Figure 2.3, which represents a three state HMM with additional entry and exit states (s_1 and s_5) typical for HMM notation in HTK[3].

In the shown example typical for ASR, transitions from state i to j with transition probabilities a_{ij} are only possible to the same or the following state. Start and end state (s_1 and s_5) are not modelled and provide the connection between two consecutive HMMs for continuous recognition, where a sequence of HMMs will be generated. The initial state probability vector for the pictured HMM is $\pi = (a_{12}, 0, 0)$ with $a_{12} = 1$ as the values of π must have a sum of one. The transition probability a_{45} from the second last to the end state is the probability that the current HMM will be left and in case of continuous recognition the next HMM will start. As we have HMMs we need to estimate the probability $b_{in} =$

[3] HTK: Hidden Markov Model Toolkit, `http://htk.eng.cam.ac.uk/`, [37]

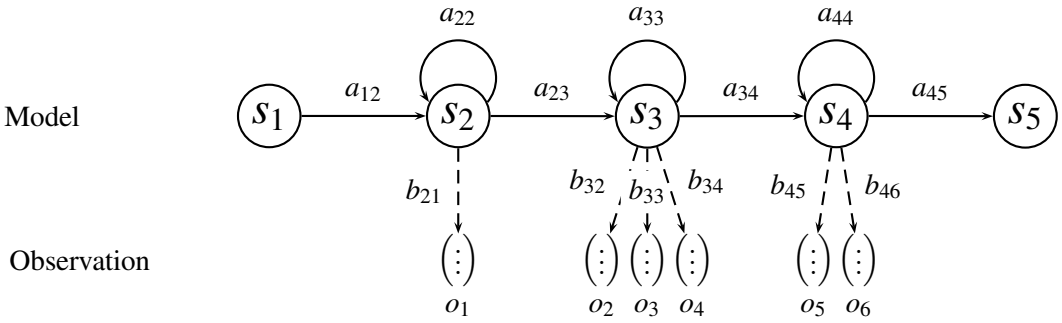


Figure 2.3: Hidden Markov model based on HTK notation ([37]). An HMM with 3 modelled states s_i plus start and end states s_1 and s_5 typical for HTK is presented. Transitions with transition probabilities a_{ij} are only possible to the same and next state. The output probability for an observation o_n given state s_i is b_{in}.

$b_i(o_n) = P(o_n|s_i)$ that a given observation o_n of an observation sequence O is the output of the state s_i. An observation o_n in ASR is described by a feature vector x_n extracted from the observed signal — in this work MFCCs as described in Section 2.2.1 are used as features. Thus, the sequence of observations O can be represented by the sequence of extracted feature vectors $X = (x_1, x_2, \ldots, x_T)$. The probability distribution of possible observations for each state as well as the transition probabilities can be learned from training data. Therefore, a Gaussian distribution is usually assumed to describe the probability distribution of possible observations for a state. As each feature vector usually has multiple dimensions, a multivariate Gaussian distribution is assumed to model the probability distribution of the observations. The mean vector μ_i and the covariance matrix Σ_i to describe the probability distribution are calculated from all observations' feature vectors in the training data that are associated with the same state s_i. During recognition, the probability that observation o_n with feature vector x_n of dimension K is output of state s_i is then calculated as follows:

$$b_{in} = b_i(x_n) = \frac{1}{(2\pi)^{K/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}\left((x_n-\mu_i)^T\Sigma_i^{-1}(x_n-\mu_i)\right)} \tag{2.9}$$

Usually, the features in the feature vector are assumed to be uncorrelated so that a diagonal covariance matrix Σ_i with non-zero values on the main diagonal only is commonly used.

Instead of single Gaussian distributions Gaussian mixtures can be used to better model the probability distribution b_{in} of rather varying observations for each state s_i. Using M Gaussian distributions $b_{i,m}$ with the weight $w_{i,m}$ for each distribution, the probability that observation feature vector x_n is output of state s_i is:

$$b_{in} = b_i(x_n) = \sum_{m=1}^{M} w_{i,m} b_{i,m}(x_n), \sum_{m=1}^{M} w_{i,m} = 1 \tag{2.10}$$

The matrix A with the transition probabilities, the matrix B with the observation probabilities and the initial state probabilities π are the parameters $\lambda = (\pi, A, B)$ defining the hidden Markov model.

Training of Acoustic Models

For training HMM-based acoustic models for ASR, speech data with a transcription of the spoken utterance is needed. Such transcription usually contains the sequence of spoken words ideally including

markers for mispronunciations, hesitations, etc. Considering the common case of a phoneme[4] or triphone[5] based speech recognition, the word transcription is mapped to a phoneme representation using a dictionary providing such a representation for each word. Each phoneme is modelled by a separate HMM. A soft assignment to which HMM λ and state of the HMM an observation belongs to, is done in a first step and the initial means $\hat{\mu}_i$ and variances $\hat{\Sigma}_i$ are calculated from the T observation feature vectors assigned to the state s_i:

$$\hat{\mu}_i = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{2.11}$$

$$\hat{\Sigma}_i = \frac{1}{T}\sum_{t=1}^{T} (x_t - \mu_i)(x_t - \mu_i)^T \tag{2.12}$$

Baum-Welch re-estimation ([42]), an expectation-maximisation (EM) algorithm, is used to recalculate the means and variances by including the probability $L_i(t)$ that observation t belongs to state i. $L_i(t)$ can be calculated by the Forward-Backward algorithm ([37, 42, 43]).

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} L_i(t) x_t}{\sum_{t=1}^{T} L_i(t)} \tag{2.13}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{T} L_i(t)(x_t - \mu_i)(x_t - \mu_i)^T}{\sum_{t=1}^{T} L_i(t)} \tag{2.14}$$

Baum-Welch re-estimation is repeated iteratively based on the re-estimated parameters until $P(O|\lambda)$ is not higher than in the previous iterations anymore. The transition probabilities are estimated in a similar way. A more detailed view on the Baum-Welch training for HMMs is given in [37, 42].

Acoustic Decoding

With transition and output probabilities a_{ij} and b_{in} of the model M, the joint probability of O to be generated by model λ for the state sequence S in Figure 2.3 is:

$$P(O,S|\lambda) = a_{01} b_1(x_1) a_{12} b_2(x_2) a_{22} b_2(x_3) ... \tag{2.15}$$

As the state sequence is usually hidden, we need to calculate the sum of all possible state sequences $S = s(1), s(2), ..., s(T)$ of model λ for observation O:

$$P(O|\lambda) = \sum_S a_{s(1)s(2)} \prod_{t=1}^{T} b_{s(t)}(x_t) a_{s(t)s(t+1)} \tag{2.16}$$

with entry state $s(1) = s_1$ and exit state $s(T+1)$ (which equals s_5 in Figure 2.3).

In practice, the maximum likelihood is usually calculated using the Viterbi algorithm ([14, 37]), a dynamic programming algorithm. Figure 2.4 visualises the Viterby algorithm, which aims at finding the best path (Viterbi path) through the shown matrix of an HMM recursively combining partial log likelihoods.

[4] Phonemes are basic units in a language which can be used to define the pronunciation of a word or utterance. By definition of the International Phonetic Association a phoneme is "the smallest segmental unit of sound employed to form meaningful contrasts between utterances". [41]

[5] Instead of modelling a single phoneme independently, triphones also cover variations in pronunciation of a phoneme dependent on the context by taking into account the previous and the following phoneme.

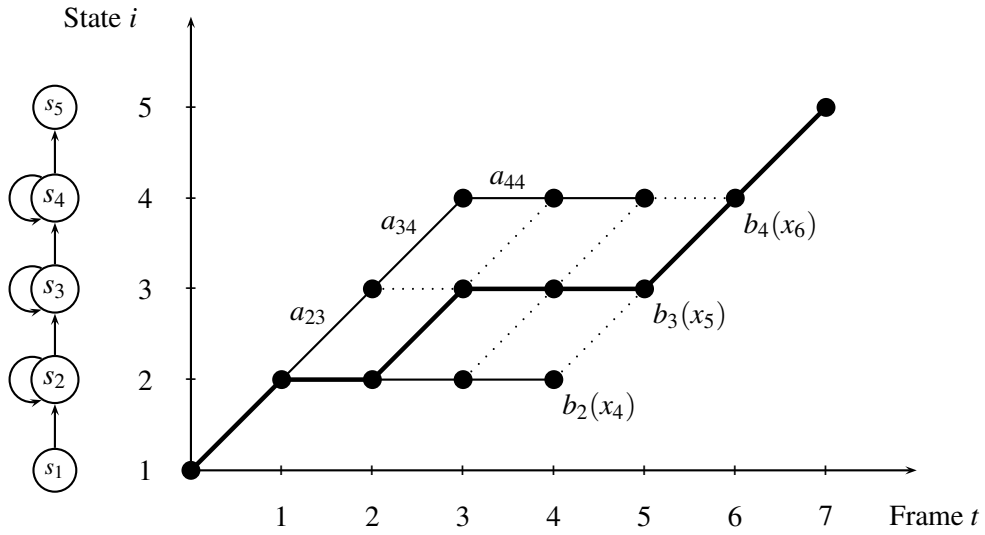


Figure 2.4: Viterbi algorithm on HMM level. For an HMM as presented in Figure 2.3 and observed feature vectors x_t all possible transitions are sketched (full and dotted lines). Several paths (full lines) are growing starting from state s_1 using Viterbi's algorithm until one path with maximum log likelihood connecting start state s_1 and end state s_5 (bold line) remains.

With $\phi_j(t)$ as the maximum likelihood for a given model λ and observation feature vectors x_1 to x_t at state j and time t, starting with $\phi_1(1) = 1$ and $\phi_j(1) = a_{1j}b_j(x_1)$, the partial likelihood can be calculated by the recursion

$$\phi_j(t) = \max_i \left\{\phi_i(t-1)a_{ij}\right\} b_j(x_t) \tag{2.17}$$

For a model with N states and T observations, we get the maximum likelihood $\hat{P}(O|\lambda)$ by calculating

$$\phi_N(T) = \max_i \{\phi_i(T)a_{iN}\} \tag{2.18}$$

Often likelihoods are replaced by log likelihoods to avoid underflow problems, so we get the maximum log likelihood by the following recursion:

$$\psi_j(t) = \max_i \left\{\psi_i(t-1) + \log(a_{ij})\right\} + \log(b_j(x_t)) \tag{2.19}$$

In Figure 2.4 we can see, how the paths are growing from the initial point $t = 0$ and $s = 1$ until the final path with the maximum log likelihood remains connecting entry and exit state.

From Word to Phoneme Models

In case of full word models in isolated word recognition, each model λ_i represents a single word w_i, and the likelihood of an uttered word is calculated as follows:

$$P(O|w_i) = P(O|\lambda_i) \tag{2.20}$$

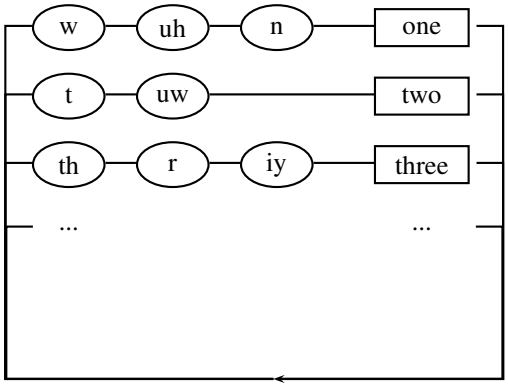


Figure 2.5: Recognition network for continuous speech recognition (adapted from [37]). In continuous speech recognition usually sub word units like phonemes are modelled by each HMM while full words are defined as sequence of such units. During recognition the sequence of sub word units with the maximum likelihood is determined and the sub word units are then replaced by the best matching words and word sequences.

In complex recognition tasks (e.g. continuous speech recognition) usually subword units (typically phonemes) are modelled instead of full words. A mapping from phoneme and triphone sequences to words is necessary. Therefore, a lexicon describes the pronunciation of each known word by sequences of sub word units introducing first lexical knowledge into the acoustic process. Figure 2.5 shows the general concept of continuous speech recognition. All words (in boxes) defined by phoneme sequences (ovals) are put into the recognition network. While an utterance is processed, phonemes and phoneme sequences are created by the acoustic models. Based on the recognition network each part of the sequence can be replaced by the best matching word in the network (usually decided on the highest log likelihood). The joint likelihood $P(O|W)$ that a word sequence W is represented by an observation O, can be calculated. Instead of using the best hypothesis only, a lattice containing the best matching words and their connections to word sequences can be created and, for example, passed to a stochastic language model (details see Section 2.2.3), where the likelihoods of the lattice can be updated accordingly.

2.2.3 Language Model

While the acoustic aspects of speech are captured by the features and described by the HMM as mentioned above, speech usually follows a certain structure (grammar). This makes certain words and word combinations more likely than others. One way to include this aspect is the use af a lexicon with word probabilities for each word and a finite grammar defining word combinations that are accepted by the system. A finite grammar is usually defined in Backus-Naur form (BNF) or similar notation grouping words into classes and defining possible structures of consecutive words. As the number of valid phrases is rather limited and requests a speaker to use exactly the specified words and structure, it is mainly used in rather restricted command and control and simple dialogue tasks.

For complex ASR tasks like large vocabulary continuous speech recognition (LVCSR) or for highly spontaneous speech the structure of the necessary grammar becomes too complex to generate, as it has to cover all aspects and exceptions. Furthermore, a speaker does not always follow a valid grammar — especially in spontaneous conversational speech — leading to recognition errors when using a finite grammar. Thus, for most more complex tasks a statistical language model (LM) is used instead. Such a language model tries to model properties of a language in a stochastic way describing the probability of

single words $P(w_n)$ and further word sequences $P(W) = P(w_1 w_2 \dots w_N)$. Usually, from a large amount of representative text data the occurrence probability $P(w_n)$ of each word w_n in the recognition lexicon is estimated in a first step. The probability of a word sequence in this simple case is then

$$P(W) = \prod_{n=1}^{N} P(w_n). \tag{2.21}$$

This is also known as unigram language model. If also word interdependencies (typical due to the language's grammar and the common usage of certain word combinations) should be considered, also conditional probabilities for certain two, three or more consecutive words (bigrams, trigrams, and in general N-grams) can be estimated and provided in a language model. The probability of a word sequence is then

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)\dots P(w_N|w_1 w_2 ... w_{N-1}). \tag{2.22}$$

Usually, the context is limited to reduce the complexity and the necessary amount of data to learn a representative language model. Furthermore, in practice the preceding words closer to the current word usually have a stronger connection to the current word. So typically bigrams (Equation 2.23) considering only the current word and the previous word or trigrams (Equation 2.24) considering the current word and the two previous words for the conditional probabilities are used:

$$P(W) = P(w_1) \prod_{n=2}^{N} P(w_n|w_{n-1}) \tag{2.23}$$

$$P(W) = P(w_1)P(w_2|w_1) \prod_{n=3}^{N} P(w_n|w_{n-2} w_{n-1}) \tag{2.24}$$

The presented N-gram approach for stochastic language modelling is still one of the most commonly used concepts in large vocabulary continuous speech recognition, even though more and more new developments in particular in the direction of neural network based language models are in the focus of more recent research of the last years ([44, 45]).

The complexity of a recognition task and the quality of the language model for this task can be determined by the perplexity of the language model. The perplexity $PP(W)$ of a word sequence of length M for a given language model is related to the cross-entropy $H(W)$ and calculated by

$$PP(W) = 2^{H(W)} \tag{2.25}$$

with

$$H(W) = -\frac{1}{M} \log_2 P(W). \tag{2.26}$$

To determine the perplexity of the language model, the perplexity of a test corpus is calculated. The perplexity can be understood as the average number of possible words, assuming an equal likelihood that can be expected at the end of each word. Thus, the lower the perplexity the better the language model is suited for the presented ASR task.

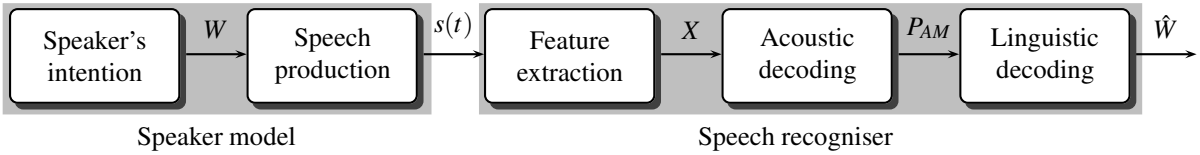


Figure 2.6: Diagram of statistical ASR work flow (adapted from [35]). A speaker has the intention to utter a word sequence W producing an acoustic speech signal $s(t)$. This signal $s(t)$ is captured and speech features X are extracted. Acoustic decoding determines the acoustic probabilities P_{AM}, linguistic decoding incorporates additional linguistic probabilities and the maximum-likelihood hypothesis for the word sequence $\tilde{W}$ is estimated.

2.2.4 Speech Decoding

During the decoding process of an ASR system both stochastic models — acoustic and language model — are combined to estimate a maximum-likelihood hypothesis $\hat{W}$ for the spoken word sequence producing observation O:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|O) \tag{2.27}$$

The a posteriori probability $P(W|O)$ is calculated with Bayes's theorem from the acoustic probability $P(O|W)$ (compare Section 2.2.2) and the language model probability $P(W)$ (compare Section 2.2.3).

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{2.28}$$

The probability of the observation $P(O)$ is constant for a given observation O and is not needed for a calculation of the maximum likelihood based on Equation 2.28. With the acoustic probabilities defined by a given set of acoustic models $P_{AM}(O|W)$ and the probabilities of word combinations defined by a language model $P_{LM}(W)$, the maximum-likelihood word sequence $\hat{W}$ for a given observation O is calculated as follows:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P_{AM}(O|W)P_{LM}(W) \tag{2.29}$$

A visualisation of this equation is shown in Figure 2.6. Still, the calculation of Equation 2.29 can be very complex even for medium sized vocabularies and complexity increases exponentially with the maximum length of the word sequence. Thus, instead of an exact calculation of $\hat{W}$ a word sequence $\tilde{W}$ is usually estimated based on methods of dynamic programming. Reducing the search space and using sequential decomposition methods lead to an estimated word sequence which is not necessarily equal to $\hat{W}$. Commonly used for estimating $\tilde{W}$ is the Viterbi algorithm analogical to the acoustic decoding described in Section 2.2.2.

2.3 Acoustic Distortions and Variability

Even though ASR performs very well in controlled environments, the development of a well performing system becomes much more difficult as soon as we have to face distortion of the signal and speech features due to the acoustic conditions and speaker influences. Several different types of speech variability and distortion influence the speech recognition performance. The two most common sources of distortion discussed and tackled in many publications on robust ASR are background noise and channel mismatch. A general model of noisy speech influenced by additive environmental noise and character-

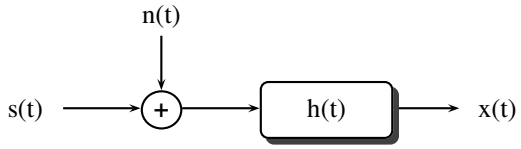


Figure 2.7: Simplified model of additive noise and channel distortion. A common model for acoustic distortion considers environmental noise $n(t)$ added to a clean speech signal $s(t)$ at the position of the microphone. Further channel effects from hardware and transmission influence this noisy signal resulting in a distorted noisy speech signal $x(t)$.

istics of hardware and transmission channel is pictured in Figure 2.7. With the transfer function of the channel $h(t)$, clean speech $s(t)$ and background noise $n(t)$, noisy speech $x(t)$ can be described by the following formula (with convolution operator $*$):

$$x(t) = [s(t) + n(t)] * h(t) \tag{2.30}$$

In reality this formulation is just a simplification as noise and speech are considered independent. Especially in very noisy environments speech is indeed influenced by environmental noise due to the so called Lombard effect, which is described in more detail in Section 2.3.2. In addition to the Lombard effect general speaker and speech variabilities already influence the speech input $s(t)$ in the system above. So exactly the same phrase spoken by two different speakers or even spoken twice by the same speaker will differ in various aspects. Finally, the process of feature extraction can already cause variations in the extracted features. In particular noise reduction and normalisation algorithms, which try to compensate or remove one of the sources of distortion above, usually introduce algorithmic distortion, which can be caused by the method itself or by erroneous estimations of compensation parameters.

In Figure 2.8 the way from the reference word sequence W the speaker plans to utter[6] to the sequence recognised by an ASR system $\hat{W}$ is shown including all major aspects of possible distortions and mismatch introduction influencing the recognition performance.

The last two blocks of linguistic and acoustic decoding are added for completeness. These blocks do not directly cause any mismatch, but they use stochastic models which provide the reference for recognition, and thus, provide the basis to which the mismatch corresponds. Generally, the better these models cover the acoustic variations and the domain of an utterance the lower the mismatch. But also for generally matching conditions, distortion can significantly influence the signal's quality and the extracted features, which generally makes an automatic recognition and discrimination of spoken units more difficult and decreases the ASR performance.

In the following we will have a closer look on main sources of distortion from acoustic environment, speech and speaker variabilities, channel characteristics towards algorithmic distortion.

2.3.1 Acoustic Environment

Distortion introduced by the speakers environment mainly includes additive noise and effects caused by the room characteristics.

[6] For practical reasons the spoken word sequence and not the planned word sequence is used as reference in ASR. Thus, the reference word sequence might contain certain effects like hesitations, mispronunciations, etc.

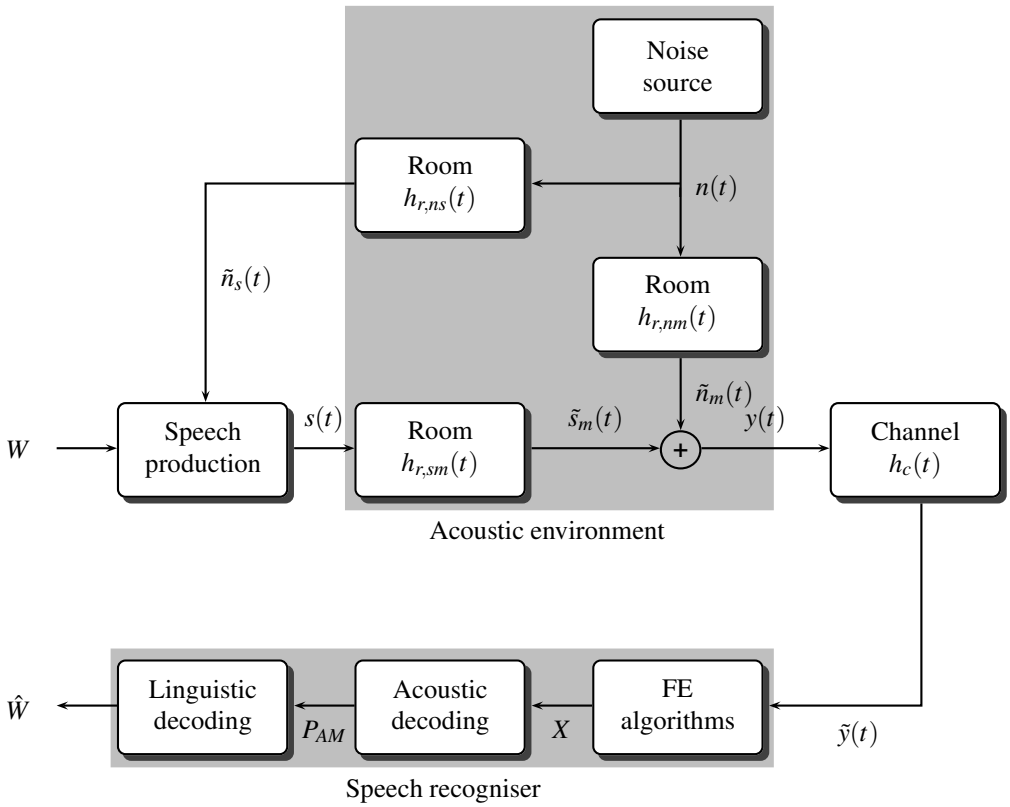


Figure 2.8: Variability and distortion in speech production and speech recognition. As opposed to the simplified model in Figure 2.7 variability and distortion of speech is much more complex in realistic conditions. Speech production is already influenced by speaker, situation as well as the acoustic environment. The influenced speech and the noise of the environment are distorted by the room characteristics dependent on their position. Channel effects (hardware, transmission, coding algorithms, etc.) and front end processing with feature extraction (FE algorithms) add further distortion to the noisy signal before the speech recogniser finally receives the extracted features for decoding.

Additive Noise

Additive noise as a source of degradation is considered in many scientific works (e.g. [46, 47]). All other environmental sound except for the desired speech is considered as additive noise also including music and other speech. For additive noise the degradation of the performance is mainly dependent on the frequency characteristics of the noise and the signal-to-noise ratio (SNR) measuring the level of speech in relation to noise. The SNR is calculated from the average powers of signal P_S and noise P_N as follows and often used in a logarithmic scale with the unit decibel (dB):

$$SNR_{dB} = 10\log_{10}\frac{P_S}{P_N} \tag{2.31}$$

For real noisy signals the SNR can only be estimated, as neither the clean speech nor the pure noise signal is known for calculation.

Many compensation techniques do not differentiate between different types of additive noise even though the influence on the speech recognition performance also depends on the noise characteristics. Generally, the more similar the characteristics of noise are to the characteristics of speech the more severe the effect on the speech recognition performance. For ASR in certain domains (e.g. broadcast data) some typical types of additive noise (e.g., background music and background speech in broadcast data — [48]) can be particularly challenging as their frequency characteristics are similar to speech and they appear much more frequent than, for example, noise with a flat spectrum. Thus, some work concentrates on such quite specific types of additive noise. [49] and [50], for example, evaluate the effect of background music on ASR or suggest a blind source separation approach to improve the speech recognition performance for speech with background music, while [51] presents an approach for separating and recognising speech of two people speaking simultaneously. Still, most noise reduction approaches try to use the general characteristic of background noise that it can be considered additive in the frequency domain (e.g. [52, 53]). As speech and noise are usually not independent (see Lombard effect in Section 2.3.2), the additive noise assumption is only a simplification and the influence of background noise on ASR is more severe and complex than often anticipated.

Room Characteristics

Another influence on both speech and noise source are the characteristics of the room. Generally, a room or environment has a transfer function changing the speech and the noise signal accordingly and thus affecting the recorded signals before they are added at the location of the microphone. The transfer function is dependent on location and frequency. In near field conditions, when a speaker is close to the microphone, the recorded speech signal is usually less affected by the acoustic environment than in far field conditions. Typical effects introduced by a room are reverberation and echoes which influence speech intelligibility and speech recognition. As reverberation and echo are major issues concerning ASR, plenty of work specialised on compensating these effects. An exemplary approach to deal with reverberant speech in ASR is presented in [54]. Effects similar to echo and reverberation can also be caused by hardware and transmission effects (Section 2.3.3).

2.3.2 Speech and Speaker Variabilities

Speaking style and speech diversity can also affect the ASR performance significantly — even in clean environments without channel distortions or additive noise. A good overview on speech variability in general and in the context of ASR is given in [55, 56]. The presented variabilities include inter-speaker variabilities. This means each speaker has a unique voice and way of speaking caused by factors

including physical differences of the vocal tract, the speaker's origin (e.g. dialects, accents), gender, etc. Common ways to reduce speaker dependent influences are vocal tract length normalisation ([57, 58]) and adaptation to a speaker by training or shifting acoustic models to the characteristics of a certain speaker (or group of speakers). But even the same speaker varies his way of speaking dependent on several factors. If speech can be planned, for example, it is rather different from unplanned spontaneous speech and speech recognition performance is affected (e.g. [48, 59]). Spontaneous speech does not only differ in the complexity and correctness of the structure of the sentence, but also in the acoustic characteristics like speed of speaking, hesitations, mispronunciations etc. Thus, the word error rate for spontaneous speech is usually higher than the word error rate for planned speech. Further variations of speech from the same speaker can be caused by changes in emotional state ([60]), sleepiness ([61]), alcohol intoxication ([61, 62]), etc.

Furthermore, the acoustic environment also influences a speaker. This effect of adapting the way of speaking to the acoustic environment and situation is called Lombard effect named after Étienne Lombard, who discovered and described the effect of certain changes in a speaker's voice in noisy environments [63]. The Lombard effect causes a speaker to change several aspects of speaking in noisy environments to make an utterance more intelligible ([64–66]). These changes include amongst others intonation, pitch, and volume. The effect is quite variable from speaker to speaker and also depends on the type and magnitude of the noise as well as the situation. Even gender and language dependent variabilities are reported in some work (e.g. [66]). [67] summarises several changes caused by the Lombard effect, which were observed in various scientific work, including amongst others:

- Increased duration of vowels, decreased duration of unvoiced sounds,
- Increase in pitch,
- Migration of energy from low and high to middle for vowels, and from low to high for unvoiced stops and fricatives,
- Increase of speech energy,
- Various effects dependent on phoneme, e.g., deletion of certain phonemes at end of word.

The Lombard effect is rather complex as we can see by the number of influences and changes. Maybe because of this complexity and variability only a very few approaches exist focusing on Lombard speech and its compensation in ASR (e.g. [68–70]).

2.3.3 Channel Effects

Various channels from microphone, hardware, signal transmission towards coding and decoding algorithms further change the recorded signal. Microphones have a certain frequency characteristic, which change the frequency characteristic of the recorded signal. In case of directional microphones this characteristic even dependent on the direction where the sound comes from. In particular alternative microphone technologies like throat microphones significantly differ in their frequency characteristics, and hence, also the features extracted from such signals differ, as we will also show in detail later in this work (Section 4.3).

In general, not only microphones but all hardware components cause distortion due to a non-ideal transfer function and other effects (e.g. [71]). Harmonic distortion, for example, adds integer multiples of existing frequencies to a signal. A common measure for the level of this distortion is Total Harmonic

Distortion (THD) defined as the ratio of the sum of the powers P_n with $n \geq 1$ of all N multiples of a frequency to the power of a frequency P_0 of the test signal:

$$\text{THD} = \sum_{n=1}^{N} P_n / P_0 \tag{2.32}$$

Harmonic distortion can be caused by amplifiers, microphones, loudspeakers and other devices. Even high quality amplifiers often have a THD of up to 1% in the relevant spectral range. While even a large THD can be acceptable in terms of perception [72], already smaller distortions change the frequency characteristics and as a result the cepstral coefficients used for ASR.

If speech is not recorded directly but transmitted via a transmission channel, the transfer function of this channel further changes the signal characteristics. Especially in digital communications, coding and decoding algorithms are often used before and after transmission. This leads to additional, and sometimes severe distortions of the signal ([73]).[7]

The introduced effects are often non-linear and can be manifold depending on the hardware and transmission channel. A compensation of general channel effects can often be achieved by normalisation approaches as briefly summarised in Section 2.4.4. Alternatively, adapted acoustic models for the specific channel can be trained to ensure a good performance of the ASR system for a particular channel.

2.3.4 Algorithmic Distortion

Algorithmic distortion can be introduced by any algorithm for manipulation of a signal. Especially in radio communications, coding and decoding algorithms used for low-bandwidth digital transmission affect the speech signal aiming at high data compression ([73]). While these are also algorithmic distortion, we consider them as part of the transmission channel and thus as part of the channel effects (Section 2.3.3).

But also after transmission several algorithms can influence the final signal and the features used for speech recognition. Interestingly, noise reduction approaches are usually able to reduce mismatch caused by additive noise, channel characteristics, etc., but at the same time also introduce new distortion. This is mainly caused by erroneous estimations of background noise or other compensation parameters causing artefacts and other effects. In the worst case the introduced noise can have more severe effects on the recognition results than the influence of the noise it tries to remove.

One example is musical noise introduced by spectral subtraction (Section 2.4.2), which is caused by an erroneous estimation of the background noise before subtraction. Thus, much work has been done to improve algorithms for spectral subtraction in such a way that the musical noise is reduced (e.g., [74, 75]). But not only spectral subtraction approaches, also practically all other robustness algorithms distort the speech signal in one way or the other. With respect to ASR the goal is to find a good balance between noise reduction and new distortion introduced to the underlying speech signal and its features.

2.4 Robustness in ASR

Due to often severe degradations of the ASR performance caused by distortion and variabilities mentioned in Section 2.3, several recommendations, techniques and algorithms have been developed and evaluated in the last decades. Two major directions are common, either trying as well as possible to

[7] We consider such coding and decoding for transmission of speech signals to be part of the transmission channel, even though they also belong to the category of algorithmic distortion in Section 2.3.4.

avoid or reduce mismatch while building an ASR system by adapting the system and providing high quality recording hardware, channel, and signal representations, or by trying to remove any mismatch during recognition applying various algorithms for robust feature extraction, mismatch compensation and online adaptation. While the robustness of a system also depends on the quality of the hardware and furthermore requires a good lexical and language model design or adaptation (as we will show in Section 4.1), we will focus on algorithms and approaches for acoustic training and adaptation as well as robust acoustic feature extraction for ASR in this chapter. Such approaches already provide a huge number of possible directions and suggestions for robust ASR brought up in the last decades.

As mentioned above robust feature extraction is one way to improve the recognition results in changing or adverse conditions. Most of these approaches try to tackle basically two major causes of mismatch: additive noise and acoustic channel mismatch (Sections 2.4.2, 2.4.4). But we can also find algorithms focused on other types of mismatch, like speech and speaker mismatch (Sections 2.4.3), or offering more general approaches for mismatch compensation (Sections 2.4.1, 2.4.6, 2.4.5). In any case, algorithms for robust feature extraction try to remove or compensate effects at some stage of the feature extraction, either on the temporal representation of the signal itself or later in the time-frequency or even cepstral domain. Other algorithms try to estimate during feature extraction or recognition, which features might be unreliable due to acoustic distortion, and either try to interpolate these features based on reliable ones or just neglect all unreliable features during recognition (Section 2.4.6).

Another way to improve recognition results in challenging conditions are offered by approaches focusing on the acoustic models. This includes advanced training, adaptation, combination, or selection of acoustic models in such a way that (in the best case) no mismatch to the extracted features exist. Especially, training or adaptation of acoustic models based on representative speech data is usually done before the recognition process when setting up the ASR system, but various approaches for online adaptation also try to adapt the acoustic models to the features and vice versa during recognition (Section 2.4.7). As properly trained or adapted acoustic models usually provide superior results compared to mismatch compensation approaches, some approaches consider multiple sets of properly adapted acoustic models with an additional model selection step to yield the best possible results for each acoustic environment in rather varying noisy conditions. Such an approach is also suggested Chapter 5 of this work, and related work in this field of research is summarised in Section 2.4.8.

In the following we will give an overview of some typical approaches of the above mentioned directions of robust speech recognition. This overview is not comprehensive but should roughly sketch the very broad research area with quite specialised fields in the context of robust ASR.

2.4.1 Feature Normalisation

Feature normalisation techniques often indirectly aim at compensation of additive noise or channel distortion. In contrast to direct compensation techniques for such distortion, feature normalisation usually makes use of statistics of the extracted features. Most common approaches include cepstral mean normalisation (CMN) or subtraction, cepstral variance normalisation (CVN), cepstral gain normalisation (CGN), and histogram equalisation.

In cepstral mean normalisation ([76]) the mean value of each cepstral coefficient $c_{n,i}$ of frame i is calculated from a long time window, for example, from a larger number of frames L, and subtracted from each cepstral coefficient:

$$c_{n,i}^{CMN} = c_{n,i} - \frac{1}{L}\sum_{\ell=1}^{L} c_{n,\ell} \tag{2.33}$$

Thus, the mean of the cepstral coefficients n within the long time window is zero. As channel characteristics in the cepstral domain are represented by additive components, cepstral mean normalisation can significantly reduce channel effects in the resulting features.

In addition to cepstral mean normalisation also the variance can be normalised by a cepstral variance normalisation ([77]). In a similar approach to CMN this approach normalises the variance of each cepstral dimension n in a long time window L with cepstral mean $\overline{c}_n$ to unity:

$$c_{n,i}^{CVN} = \frac{c_{n,i}^{CMN}}{\sqrt{1/L \sum_{l=1}^{L} (c_{n,l} - \overline{c}_n)^2}} \tag{2.34}$$

CVN mainly reduces the effect of additive noise on the cepstral features, as such noise in addition to a mean shift also affect the variance of the cepstral coefficients.

An alternative approach to CVN is cepstral gain normalisation ([78]). For each cepstral dimension n the gain instead of the variance is normalised. Therefore, the maximum and minimum values for each dimension are estimated to normalise the gain to unity:

$$c_{n,i}^{CGN} = \frac{c_{n,i}^{CMN}}{\max_l (c_{n,l}) - \min_l (c_{n,l})} \tag{2.35}$$

The effect of compensating the impact of additive noise on ASR performance is similar to CVN, but the results are reported to be superior.

Another statistical approach to feature normalisation is histogram equalisation ([79]). For MFCC-based feature extraction, histogram equalisation can be implemented in different stages of the extraction process. Often, it is implemented after the mel-filtering and before the cepstral decorrelation. Generally, the cumulative density function (CDF) of the values is determined and the values are transformed by a (possibly non-linear) function that minimises the Kullback-Leibler divergence between the density of the transformed data and the density of the training data. So the transformed features will have a CDF closer to the CDF of the training data. In quantile based histogram equalisation ([80]), only a few quantiles of the CDF instead of the full information are used to determine the transformation to adapt the data and the CDF. In that way quantile based histogram equalisation is much faster than normal histogram equalisation providing a comparable recognition performance.

All of these feature normalisation techniques, especially CMN, are widely used. As many different types of distortion — in particular additive noise and channel distortion — also influence the mean and magnitudes of the cepstral values, small to medium improvements in ASR accuracy are obtained in many situations of mismatched conditions when applying such normalisation techniques. Even in matched conditions small improvements can be obtained by reducing the effect of common variations in speech.

2.4.2 Additive Noise Reduction

Additive noise is usually removed or reduced making use of the assumed additive characteristics by trying to estimate the noise spectrum to subtract it from a noisy speech signal. These approaches are called noise reduction approaches and mainly differ in the way of noise estimation and the algorithm used for removing noise trying to avoid distortion introduced by incorrect estimations.

One of the most widely used concepts for noise reduction is spectral subtraction strongly based on the assumption that background noise and speech are completely independent from each other so that a noisy signal can be described by adding the pure noise signal to the pure clean speech signal. This

concept for noise reduction was introduced in [52] in 1979 and influenced several algorithms for reducing background noise in the last decades. The basic idea is that additive noise — as it is assumed to be uncorrelated to the clean speech signal — can easily be removed by estimating the spectrum of the noise signal and subtracting this spectrum from the noisy speech spectrum of the signal. Usually, the noise characteristics are estimated from non-speech parts of the audio signal, which further assumes that the noise is quasi stationary. In spectral subtraction the phase-less (estimated) spectrum of clean speech $|\hat{X}(f)|$ is estimated from the noisy speech spectrum $|Y(f)|$ and time-averaged noise spectrum $|\overline{N(f)}|$:

$$|\hat{X}(f)|^{\beta} = |Y(f)|^{\beta} - \alpha|\overline{N(f)}|^{\beta} \tag{2.36}$$

The factor α determines the amount of noise to be subtracted (with $\alpha = 1$ for full subtraction). β is either 1 for magnitude spectral subtraction or 2 for power spectral subtraction. The noise signal $N(f)$ is not known and must be estimated from the noisy speech signal, usually using the spectrum of non-speech frames known from voice activity detection. Unfortunately, erroneous estimations of the background noise cause distortion and so called musical noise typical for spectral subtraction. Thus, much work since the introduction of the approach has been done to improve algorithms in a way that musical noise is reduced (e.g., [53, 74, 75]).

Wiener filter approaches are widely used for robust speech recognition, especially for improved additive noise reduction. The fundamentals of Wiener filtering were introduced by Norbert Wiener in 1949 ([81]). The Wiener filter is optimal in the sense of least mean squared distance between filter output and reference signal. A straightforward approach to Wiener filter noise reduction is closely related to spectral subtraction. Equation 2.36 with $\beta = 2$ and $\alpha = 1$ can also be written as

$$|\hat{X}(f)|^2 = H(f)|Y(f)|^2 \tag{2.37}$$

with

$$H(f) = 1 - \frac{|\overline{N(f)}|^2}{|Y(f)|^2} = \frac{|Y(f)|^2 - |\overline{N(f)}|^2}{|Y(f)|^2}. \tag{2.38}$$

$H(f)$ is in the range of 0 to 1 and is an SNR-dependent attenuator. Comparing Equation 2.38 with the frequency response of a Wiener filter for noise removal $W(f)$ in the following Equation 2.39 with Expectation $E[.]$ (compare [75]) shows the analogy between both concepts.

$$W(f) = \frac{E[|Y(f)|^2] - E[|N(f)|^2]}{E[|Y(f)^2]} \tag{2.39}$$

Instead of the instantaneous noisy speech spectrum and the time-averaged noise spectrum, the expectation (*ensemble-average*) of each spectrum is used. The Wiener filter coefficients with a value between 0 and 1 (dependent on the signal-to-noise-ratio (SNR)) attenuate the signal the more the more noisy the speech signal is. Wiener filtering approaches usually show good results in improving the ASR results for speech with additive noise and are widely used. More information about general aspects of Wiener filtering and about Wiener filtering for noise reduction can be found in [29, 82].

Spectral subtraction and especially the Wiener filter approach are commonly used when only additive noise is expected. They are usually effective in reducing additive influences, in particular for medium SNRs, where a voice activity detection for a noise spectrum estimation provides reliable results. Other common sources of distortion are not considered at all by such approaches, so that additional compensation steps must be considered in case of scenarios also including other types of distortion. This includes the Lombard effect, which cannot not be compensated by these approaches but that typically

occurs in conjunction with loud additive background noise. Both algorithms can introduce considerable algorithmic distortion mainly caused by erroneous noise estimations. Thus, for a good performance in terms of ASR accuracy it is recommended to train the acoustic models already on features processed by these algorithms to reduce the influence of such algorithmic distortion during recognition. Especially the Wiener filter approach related to spectral subtraction is widely used for additive noise reduction as this optimal filter in mean squared sense hardly affects clean speech but significantly reduces influences of additive noise on the signal as long as a good estimation of the voice activity is provided.

2.4.3 Speech and Speaker Mismatch Compensation

Speech and speaker variation is rather complex and often difficult to compensate. Practically, speaker adapted acoustic models are often used, ideally covering roughly the speaking style also typical during recognition. Typical model adaptation techniques are MAP and MLLR as detailed in Section 2.4.7. In the following we will focus on two of the various aspects of speech variability: speaker normalisation reducing effects of inter-speaker variations and approaches for Lombard effect compensation.

Speaker Normalisation

A common approach of speaker normalisation is vocal tract length normalisation (VTLN, [57, 58]). The idea behind VTLN is the compensation of the effect of the vocal tracts on speech production. Generally, the vocal tracts of different speakers have a different length influencing mainly the fundamental frequency and thus the formant frequencies of speech. To compensate for this effect, the frequency axis can be warped by a factor α dependent on a speaker:

$$F_{VTLN} = \alpha F \tag{2.40}$$

Several approaches for estimating factor α exist. A common approach is a model based maximum-likelihood approach by Lee and Rose [58]). During training the training data is split into two sets with different speakers. One set is used for training initial HMMs. Then the utterances of each speaker in the other set are Viterbi aligned using the HMMs, and the factor α for each speaker is estimated based on all utterances of this speaker. The best warping factor α for each speaker is determined by applying various different factors to his utterances and deciding for the one maximising the likelihood when decoding on the trained HMMs. The procedure is repeated after swapping both sets for training and estimation using the normalised utterances of the speakers for training of the HMMs. This process is iterated until all α for the speakers do not change significantly any more. A set of acoustic models is finally trained from all training data warped by their respective factor.

During recognition a first recognition hypothesis is generated from the unwarped utterance. After alignment of the HMMs based on this hypothesis, several warping factors are applied and the α maximising the likelihood of the warped utterance for the aligned HMMs is applied to decode the utterance. Certain variations of VTLN exist trying to improve the normalisation or increase the processing speed (e.g. [83, 84]).

VTLN successfully proved to reduce mismatch caused by one of the major sources of inter-speaker variations, the vocal tract. It is often applied when processing time is not a critical requirement. While VTLN yields small to significant improvements in ASR accuracy dependent on the speakers characteristics, the estimation of the warping factor α during recognition by using a maximum-likelihood approach is computationally expensive. Even though approaches for faster processing were researched in the last years, this is still the most critical aspect when considering implementation of VTLN.

Lombard Effect Compensation

Due to the dynamics of the Lombard effect, a compensation is rather difficult. One approach of a compensation of the cepstral features for robust ASR was introduced by [68]. The presented approach tested on Korean speech is based on a degradation model covering various aspects of noisy Lombard speech including both Lombard effect and noise contamination. They show how Lombard effect and additive noise are reflected in the cepstral features, which indicates that an approach of multi-linear regression can help for compensating both variabilities. With c_n^{clean} as the n^{th} clean speech cepstral coefficient and $c_k^{Lombard}$ as the k^{th} noisy Lombard speech cepstral coefficient for an aligned frame representing the same state, multi-linear regression can be used to estimate transformation matrix A and vector b, which solve the following equation with minimal error:

$$c_n^{clean} = \sum_{k=0}^{K} A_{n,k} c_k^{Lombard} + b_n \tag{2.41}$$

While the results for such a compensation reported in [68] show significant improvements, similar experiments on throat microphone data of British English speech recorded in a noisy environment could not show any clear improvements compared to the baseline system ([69]).

More recent work in Lombard effect compensation is summarised in [70]. Methods for unsupervised spectral and cepstral equalisation are proposed and combined with a codebook based selection of the best matching noisy models.

The first normalisation using warp and shift frequency transform is based on VTLN. While VTLN only allows a warping of the frequency axis by a factor α, an additional additive term β is introduced for warp and shift frequency transform:

$$F_{WS} = \alpha F + \beta \tag{2.42}$$

This extension to warp and shift allows to a certain level different shift rates and directions for lower and higher formants compared to VTLN. This promises to improve Lombard effect compensation, where a shift of formants can be more complex than anticipated for general formant shifts from speaker to speaker. During recognition the best parameters from a parameter search grid are determined per utterance maximising the decoding likelihood.

A further normalisation step after warp and shift in this approach is quantile based cepstral dynamics normalisation (QCN). Instead of normalising on the minimum and maximum as done in CGN (Section 2.4.1), all samples i of a cepstral value $c_{n,i}$ are sorted and the cepstral values are normalised on the low and high quantile $q_j^{(c_n)}$ and $q_{100-j}^{(c_n)}$ at position $round(jL/100)$ and $round((100-j)L/100)$ in a sorted list for the L frames:

$$c_{n,i}^{QCN_j} = \frac{c_{n,i} - \left(q_j^{(c_n)} + q_{100-j}^{(c_n)}\right)/2}{q_{100-j}^{(c_n)} - q_j^{(c_n)}} \tag{2.43}$$

QCN compared to CGN is supposed to have the advantage that it is less sensitive to outliers as it avoids to normalise on the two extreme values.

In addition to the normalisation techniques the authors also provide several sets of HMMs trained on artificially noisy speech of different SNRs. During recognition the normalised features are used to decode the utterance with each of these sets of HMMs. From the best hypotheses of all sets of HMMs the one with the highest likelihood is taken. The authors report improvements compared to standard normalisation techniques on noisy Czech Lombard speech.

It can be expected that in the presented approaches for Lombard effect compensation the general normalisation of noise, channel and speaker effects make up for the major improvements reported in the related publications. While these approaches focus on the Lombard effect in particular, they both apply methods closely related to general speech and speaker adaptation as well as noise and channel compensation. Considering the highly dynamic characteristics of the Lombard effect as discussed in Section 2.3.2, the presented approaches more likely present variants of successful normalisation and adaptation techniques than particular Lombard effect compensation algorithms. Still, in particular the extended VTLN approach and the quantile-based cepstral normalisation proposed in the latter work are promising adaptations of existing normalisation techniques.

2.4.4 Channel Compensation

One of the most common approaches applied to reduce certain channel effects is cepstral mean normalisation (CMN) as described in more detail in the general feature normalisation section (Section 2.4.1). CMN is actually quite powerful as one of the reasons and advantages of using cepstral coefficients is the deconvolution of channel effects and speech characteristics, so that both components result in additive terms in the cepstral domain. Assuming constant channel characteristics, the subtraction of the cepstral mean from each cepstral coefficient will successfully remove the channel effects without affecting the speech features — as long as the speech features can be assumed to be centred. If we already apply CMN during training, this requirement is approximately fulfilled. Still during recognition only an estimate of the mean can be calculated for an utterance, as especially for short utterances the cepstral features of speech of the particular utterance might not be centred. Furthermore, including non-speech frames for cepstral normalisation can affect the centring of speech frames, especially in case of environmental noise.

This aspect is, for example, addressed in [85]. The authors suggest to generalise CMN to two different classes: speech and noise. During training the mean of all cepstral vectors tagged as noise μ_n^{train} or speech μ_s^{train} can be calculated. As discrimination of speech and noise frames is not working perfectly, an a posteriori probability p_i of frame i being noise is estimated during recognition. The mean μ_i to subtract from a frame i is then calculated from these values and the cepstral noise and speech means of the utterance μ_n^{rec} and μ_s^{rec} by the following equation:

$$\mu_i = p_i(\mu_n^{train} - \mu_n^{rec}) + (1 - p_i)(\mu_s^{train} - \mu_s^{rec}) \tag{2.44}$$

The estimated mean vector for a frame μ_i is then subtracted from each frame's cepstral vector c_i to estimate the normalised vector $\hat{c}_i = c_i - \mu_i$. An additional 25% and 5% relative improvement in word error rate compared to classical CMN on mismatched and matched conditions is reported for this approach named Augmented CMN.

While many more approaches for channel compensation exist, classical CMN or one of its variations is commonly used in state-of-the-art speech recognisers for more than a decade because of its efficiency and low complexity. Augmented CMN increases the complexity of CMN and requires a voice activity detection with soft decision also providing the a posteriori probability for each frame. The complexity is still relatively low and a separation of the noise and speech mean is usually relevant. Nowadays, CMN is often applied including such a weighting scheme or considering speech frames only.

2.4.5 Robust Feature Extraction

Considering the different approaches dealing with different sources of distortion, it can be a good idea to combine certain approaches to improve the recognition performance of realistic speech data, which

usually suffers from more than one source of distortion.

An approach standardised by the European Telecommunications Standards Institute (ETSI) is the ETSI advanced front end ETSI ES 202 050 ([86]). The extraction of robust speech features for ASR is just a part of the standard, which also covers aspects of distributed speech recognition like compression, transmission and quality aspects. The feature extraction of ETSI ES 202 050 contains four stages: noise reduction, waveform processing, cepstrum calculation, and blind equalisation. Noise reduction is based on a two stages Wiener filter. First, the signal is generally de-noised by estimating the spectrum of noise from pure noise segments determined by voice activity detection. Second, SNR-dependent noise reduction by dynamic Wiener filtering is applied. The next stage is a SNR-dependent waveform processing of the noise reduction output before the cepstrum of the processed waveform is calculated. For cepstrum calculation pre-emphasis is performed before the cepstral coefficients are calculated analogously to Section 2.2.1. Furthermore, the log energy is calculated. The cepstral calculation outputs feature vectors with 14 dimensions, the log energy coefficient and the first 13 cepstral coefficients $c(0)$ to $c(12)$. The last step of the feature extraction is blind equalisation of the coefficients based on a least mean squares algorithm as detailed in [86]. On server side (after transmitting the features) log energy and $c(0)$ are joint and the derivatives of first and second order of this value and the twelve cepstral coefficients $c(1)$ to $c(12)$ are calculated resulting in 39 dimensional feature vectors used for ASR.

In more recent research extensions and modifications of the ETSI advanced feature extraction can be found. The authors in [87], for example, propose a way to reduce the complexity of the ETSI advanced front end, which mainly aims at enabling the ETSI advanced front end to also operate on mobile devices with low computational power.

Another front end approach for robust feature extraction is described by Segura et al. in [88]. This approach uses spectral noise subtraction (compare Section 2.4.2) to reduce the effect of additive noise and a complementary cepstral histogram equalisation (compare Section 2.4.1). The authors report very good results on the Aurora 2 dataset and SpeechDat Car (Finnish, Spanish, and German set) emphasising that both approaches show significant improvements on the recognition rate — each approach on its own and especially both approaches together included in a common work flow. The complete work flow provides an average of 7 to 14% absolute improvement in word accuracy for the SpeechDat Car sets and an average of 4.5 and 25% absolute improvements in case of matched and mismatched conditions for the Aurora 2 evaluation sets.

In general, a combination of additive noise reduction and feature normalisation techniques as applied in the presented robust feature extraction approaches should be considered in most real-life situations where both background noise and channel variations can be expected. The techniques applied in the ETSI advanced front end and the approach by Segura et al. had been proven to be successful on their own, and thus, promise to provide robust front ends for feature extraction and normalisation. This is in line with the reported results for both approaches. While these approaches are usually successful in mismatched conditions, they pose the problem that algorithmic distortion can slightly reduce the ASR performance in clean conditions. Thus, their application is mainly recommended when noise and acoustic distortion is expected in a larger scale during recognition.

2.4.6 Missing Feature based Approaches

Missing feature based approaches assume that the features extracted for ASR suffer from distortion, but that this distortion is rather different in magnitude from feature to feature in most cases. Thus, assuming that an estimation of reliable and unreliable features is possible, only the reliable features can be used for further processing using these features to interpolate or restore the other “missing” features in one way or another. Missing feature approaches include a rather large group of algorithms.

Most missing feature based approaches work in some time-frequency domain usually after mel-filtering the spectrogram of the windowed signal. The first and most difficult step is to detect unreliable features in this spectrogram. Various approaches are used in missing feature approaches for this purpose. A common approach estimates the signal-to-noise ratio (SNR) for each spectral component and decides based on a threshold, whether the feature is considered to be reliable or not. But also other principles can be used to make this decision. In a so called spectrographic mask, the reliable and unreliable features can be marked — either using a hard decision classifying the features or by soft masking. After the features are classified, basically one of two general approaches can be chosen to deal with unreliable features:

- *Feature-vector imputation*: unreliable features are reconstructed — here a variety of approaches exist — and a full set of features is used in a standard classifier. Two major approaches are commonly used, *correlation-based reconstruction* and *cluster-based reconstruction*. The first tries to reconstruct the missing features from reliable components in the neighbourhood, the latter from the distribution of the cluster.

- *Classifier modification*: the classifier is modified to be able to interpret both reliable and unreliable features. Either the classifier completely ignores unreliable components (*Marginalisation*) or it tries to estimate improved features for the unreliable parts from the reliable ones (*Data Imputation*).

More details on the concept of missing features and the most common approaches are given in [28].

Missing feature based approaches showed to be efficient for ASR in noisy conditions, in particular for additive noise mainly affecting only parts of the spectro-temporal representation of a speech signal. The most difficult aspect is the extraction of the spectrographic mask that defines the reliable and unreliable areas in the spectrogram. A drawback of most missing feature approaches — especially for classifier modification methods — is the application in the spectral domain often requiring spectral features instead of cepstral coefficients. Hence, an increased ASR performance is achieved compared to the baseline using spectral features, but improvements are often much smaller compared to a setup using cepstral coefficients with a generally better performance.

2.4.7 Model and Feature Adaptation

The most common model adaptation techniques in ASR are Maximum A Posteriori (MAP - [37, 89, 90]) and Maximum Likelihood Linear Regression (MLLR) adaptation ([23, 37, 91]). While these approaches are mostly applied to achieve speaker adaptation, they are not limited to this use case. Both approaches use a small amount of data from the speaker or acoustic environment to adapt to, to shift the acoustic models trained on general data towards the new data characteristics.

In MAP adaptation prior knowledge about the model parameter distributions is used. This prior knowledge is, for example in case of speaker adaptation, the general speaker independent model. Given the mean vector of the general model μ_{jm} and the mean vector of the adaptation data $\overline{\mu}_{jm}$ of state j and mixture m (of a HMM with Gaussian mixtures) with weighting τ of a priori knowledge and occupation likelihood N_{jm} of the adaptation data, the adapted mean is calculated by

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\overline{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm} \tag{2.45}$$

The occupation likelihood N_{jm} is the sum of the likelihoods $L_{jm}^{(r)}(t)$ for all observations R and times of an observation T_r:

$$N_{jm} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^{(r)}(t) \tag{2.46}$$

From all observations' feature vectors $x_t^{(r)}$ weighted by the likelihoods the mean vector for state i and mixture m of the adaptation data is calculated:

$$\overline{\mu}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^{(r)}(t) x_t^{(r)}}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^{(r)}(t)}. \tag{2.47}$$

The new values are stored in a new set of speaker or environment adapted acoustic models. As MAP is working on component level updating each component separately, it requires a rather large amount of adaptation data.

The MLLR approach for adaptation has much lower requirements concerning the amount of adaptation data. MLLR estimates a set of linear transformations that reduce the mismatch between the general model and the adaptation data. For that purpose the means and variances of the Gaussian distributions are transformed. The adapted mean vectors $\hat{\mu}$ can be calculated from the general mean vector μ of dimension N by the transformation with an $N \times N$ transformation matrix A and bias vector b weighted by factor w (often set to 1):

$$\hat{\mu} = A\mu + bw \tag{2.48}$$

This equation is also often formulated as $\hat{\mu} = W\xi$ with A and b combined in a single matrix $W = [bA]$ and extended mean vector $\xi = [w\,\mu_1\,\mu_2\,\ldots\,\mu_N]^T$. In case of model adaptation the variances Σ of the models are also transformed by a $N \times N$ covariance tranformation matrix H:

$$\hat{\Sigma} = H\Sigma H \tag{2.49}$$

The matrices for transformation are estimated by Expectation-Maximisation (EM). As MLLR does not adapt each component separately but determines a general transformation, it requires less adaptation data than MAP.

MLLR can also be used to adapt the features instead of the Gaussian distributions with means and variances. This approach is commonly referred to as Constrained MLLR (CMLLR). Instead of the means of the Gaussian distributions the observation feature vectors x are transformed by $\hat{x} = Ax + bw$ equivalent to Equation 2.48.

In [92] an approach of factored transforms based on CMLLR to estimate and compensate speaker and environmental variabilities in separate steps iteratively is presented. In contrast to the conventional approach, the authors assume that a separate linear transform exists for environmental variability $W_e = \{A_e, b_e\}$ and for speaker variability $W_s = \{A_s, b_s\}$, which leads to the following equation

$$\hat{x} = A_s(A_e x + b_e) + b_s \tag{2.50}$$

This equation is equivalent to $\hat{x} = A'x + b'$ with $A' = A_s A_e$ and $b' = A_s b_e + b_s$. Instead of estimating a general A and b, environmental and speaker variabilities are estimated alternately with the respective other parameters fixed from appropriate training data. During recognition a transform can be applied either for both variabilities at the same time or for each variability separately dependent on the situation making this approach more flexible than standard CMLLR. The approach can also be incorporated into acoustic model adaptation strategies as shown in the publication.

MAP and MLLR are standard approaches for acoustic model adaptation, especially speaker adaptation, widely used today whenever the amount of training data is too little to train dedicated acoustic models. Improvements compared to not adapted acoustic models depend on the quality of the original model and the adaptation data as well as the mismatch between both. In many cases significant improvements are achieved. CMLLR and related approaches are promising approaches for online adaptation during recognition. Due to a multi-linear transformation as opposed to statistical normalisation approaches more complex patterns of mismatch can be successfully compensated by such adaptation techniques. Drawback is the requirement of sufficient speech samples of each speaker or acoustic condition to reliably estimate the respective transformation matrices.

2.4.8 Multi-Model Approaches

Well trained or adapted acoustic models usually yield the best results in ASR as ideally no acoustic mismatch is present between training data and data to be recognised. Thus, some approaches are based on several sets of well trained acoustic models, each specialised for a certain acoustic environment. One group of approaches consider using all sets of acoustic models in parallel to produce a seperate hypothesis for each of them in a first step, fusing all or the N-best hypotheses in a second step to yield the final result. Another group of approaches includes an additional acoustic model selection step before the main ASR process using only the best suited set of acoustic models for recognition. The second approach is often faster as a full recognition process is only performed for one set of acoustic models, while the first approach requires a full recognition process for all of the sets.

In [93] an approach for acoustic model selection is proposed that uses the silence model of each set of acoustic models to determine the best suited acoustic models based on the additive noise characteristics. In a first step, the first few frames of a speech segment (which usually contain noise only) are used to find the acoustic models with the highest probability based on the silence models.

During initialisation the noise data (without speech) is clustered using k-means based on the similarity of their Gaussian distribution density functions of each cepstral coefficient. As artificial noisy speech data is used in the experiments (Aurora 2, [47]), this clustering is rather straightforward compared to realistic conditions, as the noise data is available separately. The noise from each cluster is then added with an SNR of 5dB to the clean speech training utterances. For each of the noise clusters an individual set of HMMs is trained from this noisy speech data. The background noise of each set is modelled by a single state model with a mixture of 8 Gaussian distributions. During recognition the single state background models of each set of HMMs are used to calculate the generation probability of the background noise taken from the first frames of the utterance. The set which provides the highest probability is then chosen for recognition.

The performance of the model selection process of this approach is not directly quantified in the paper, but the presented results of the noise classification roughly match the subjective expectations. Thus, the information of the first frames seem to be sufficient to determine the correct background noise in most cases. A major advantage of the approach is that it is very fast, as a matching of the silence models hardly takes any time. But it is limited to additive background noise which must already be present at the beginning of an utterance. As no speech information is considered for classification, factors like speech diversity, channel characteristics and other effects can hardly be covered by such an approach of acoustic model selection.

A similar approach, but with separate noise models and additional signal-to-noise-ratio classification was introduced in [94]. Instead of using the silence models trained for each set of acoustic models, separate Gaussian mixture models (GMMs) with four mixtures were trained on the cepstral coefficients for each class of noise on characteristic noise samples. Practically, this is identical to the one state noise

model within the sets of HMMs in [93] — except for a different number of mixtures. As opposed to [93] the noise is not clustered automatically, but the given Aurora 2 noise sets are used as provided to train acoustic models for each noise type and for different SNRs. Therefore, each type of noise is added to clean speech with SNRs from 0dB to 30dB in steps of 2dB and a model is trained for each type and SNR. Noise classification is then performed using the first 10 (non-speech) frames of each utterance. Additionally, the SNRs for each utterance are estimated with a simple voice activity detection (VAD). Both, the SNR information and the result of the noise classification are used to determine the set of acoustic models for the recognition process. Interestingly, for low SNRs of the test data, the best performance of ASR is reported for the HMMs trained on a set with higher SNRs than the test set. This means that a match in noise type and SNR does not necessarily yield the best results. In general, the authors recommend to use 3-5 different sets of HMMs per noise type (ideally for SNRs of 5dB, 10dB and 20dB) as a higher number hardly provides any improvement but increases complexity. Using a setup of 3 sets of models for each of the 4 noise types, the proposed method shows good results on the Aurora 2 dataset, especially when incorporating an additional mismatch compensation step. Again, the approach is limited to additive noise only.

In [95] the authors present a multi-model approach that tries to improve the results by considering speaker variabilities. In a first step they pre-cluster all speakers in the training data based on their acoustic similarity measured by Gaussian log likelihood. For each of the resulting clusters of similar speakers a separate set of acoustic models is trained. As the number of training utterances is rather low in this work, general speaker independent acoustic models are adapted with the cluster's training data using maximum a posteriori (MAP) adaptation. During speech recognition a cluster selection is introduced to determine the top N clusters closest to the acoustic features using Euclidean distance. The distance is calculated after a first speech recognition pass with a speaker independent set of acoustic models and Viterbi-alignment against the ASR transcription to be able to calculate the distance of each frame to the aligned observation of the cluster-dependent models. In a next step maximum likelihood linear regression (MLLR) is applied to adapt the cluster-dependent models to be closer to the test speaker's features. The reported results on the Wall Street Journal database for the presented clustering, model selection and adaptation approach showed significant improvements by NIST definition compared to MLLR adaptation on a speaker-independent model (26.5% relative improvement compared to 22.5%). In their conclusion the authors suggest that the phonetic information they used for speaker clustering could possibly be ignored and vector quantisation based information might be used instead to characterise a speaker. This proposed method provides good results for multi speaker situations, but requires a full speech recognition pass for each of the provided sets of acoustic models for a subsequent alignment. Thus, the approach is not very fast. Even though an extension to general acoustic clustering and application of the concept might be possible, the authors only focus on and evaluate the aspect of speaker modelling and speaker adaptation.

In [96], [97] and [98] tree-based model selection approaches are presented that aim at a selection of the best matching set of noise or speaker-dependent models for a new utterance presented to the system.

The authors in [96] and [97] follow the same concept of tree-based model selection, while in the first work different noise conditions and in the second work different speaker characteristics are considered. In tree-structured noisy speech modelling various acoustic model sets are trained based on top-down clustering. In each of the leaf nodes a set of acoustic models for a single noise type at a single SNR is provided. In each step getting closer to the root node more and more SNRs and/or noise types are clustered by their similarity and a set of models for each of the clusters is provided. The root node finally contains a set of acoustic models trained on all noisy speech data. The authors train an additional GMM classification model for each of the nodes, which contain a specialised set of acoustic models for recognition. In case of tree-structured speaker modelling the same concept is adapted to speaker-dependent

models. This means that in each of the leaf nodes single speaker HMMs are provided, while the root contains a set of general multi-speaker models. In both works the similarity measure and clustering algorithms are not detailed. During recognition of a new utterance the tree is traced from the root to the nodes estimating the best matching set of models based on the classification model providing the highest likelihood. In both publications a MLLR adaptation approach for the mean values is applied to improve robustness in mismatched conditions. Improvements compared to baseline and multi-conditional acoustic models are reported. The application of the presented tree-based approaches for two different types of acoustic variabilities indicate that such approaches could possibly be generalised to consider acoustic variabilities beyond speaker and additive noise. Still, both approaches are designed and evaluated for one aspect only. Furthermore, the training data of each set of acoustic models is needed to train dedicated GMMs for each node.

In [98] an approach of acoustic model selection based on a generated speaker tree is described. The authors use an agglomerative bottom-up approach to create the tree. In an initial stage each speaker is clustered in a separate node and a GMM with two mixtures is trained on the speech frames of the node's speech data to model the node. In the next step the Bayesian Information Criterion (BIC - [99]) is used as measure to decide which clusters are most similar to each other and will be merged. For this new cluster a new GMM is trained with a total number of mixtures equal to the sum of mixtures of the two merged nodes. This process is repeated until the top node containing all speakers is reached. Due to this generation process the resulting tree can be unbalanced. For model selection a two-step approach is used to find the best matching model to a speaker in this tree. First, the tree is pruned to the best matching single branch for the speaker following the maximum-likelihood path from the root to the leaves. Then, the best matching set of models from this branch is selected decided on the likelihood scores of each GMM. For evaluation the authors used a corpus of disordered speech (Alborada-I3A - [100]) with 232 unimpaired speakers for training and 14 impaired speakers for recognition. Two cases were evaluated. First, the best matching selected acoustic models of the tree were used for recognition. Second, MAP adaptation of the best matching selected acoustic models of the tree with part of the test data of impaired speech was performed and the adapted models were used for recognition. In both cases small improvements in terms of word error rate compared to the results when using the root models containing all speakers were reported. The authors argued that one reason for the rather small improvements might have been that the speakers in the training set might be rather similar, as only children and young adults are part of the set. This bottom-up approach of a tree-based model selection is also focused on speaker variabilities only. Furthermore, the evaluation of acoustic models trained on unimpaired speakers and tested on impaired speakers makes it difficult to compare and assess this approach. We expect a similar performance in case of unimpaired test speakers as in [97] with a similar concept but top-down clustering of the speakers presented before.

Generally, acoustic model selection approaches using separate acoustic models for different speakers or noise conditions report small to significant improvements compared to the baseline results for multi-conditional acoustic models. The presented approaches are often much faster than a parallel decoding for each set of acoustic models and a fusion of the resulting ASR hypotheses, which is another option of incorporating multiple sets of acoustic models. Most of these approaches focus on one aspect of speech variability and distortion only — usually speaker variations or additive background noise. This significantly limits the performance and the field of application in real-life situations. We believe that a multi-model approach not limited to a certain source of variability and distortion is possible. Such an approach is proposed and evaluated in Chapter 5.

2.5 Evaluation of ASR Systems

ASR systems are evaluated on reference data offering a transcription of the ground truth of the spoken content. This test data should preferably not be part of the training data. Otherwise, the recognition performance can change as the utterances are known to the system. Various evaluation corpora offering distinct training and test sets exist (e.g. [1, 47, 48]) providing different levels of complexity in terms of task (from digit recognition to LVCSR) and acoustics (from clean speech to different levels of distortion and mismatch). As a measure of performance of ASR systems the word error rate and related measures as detailed in the following section are used.

2.5.1 Evaluation Measure: Word Error Rate

In ASR the quality of the recognition output can be measured by comparing the results with the reference transcript of test utterances. Considering a recognition on word level, the word error rate (WER) and the word accuracy rate (WAR) are common measures. Relevant for calculating the quality of an automatic transcript are the number of correctly recognised words and the number of recognition errors, which can be substitutions, deletions or insertions:

1. Correct words **C**: Words that are identical in the recognition result and the reference transcription.
2. Substitutions **S**: Words in the reference transcription that are replaced by another (often acoustically similar) word during recognition.
3. Deletions **D**: Words that are missing in the recognition result.
4. Insertions **I**: Words that appear in the recognition result but not in the reference transcript and do not replace another word.

As there is usually not just one unique way how an erroneous recognised word can be mapped to a reference word, Levenshtein distance is used to determine the number of each of these recognition errors. Levenshtein distance is named after Vladimir Levenshtein, who introduced this string metric in 1966 ([101]). It is the minimum number of edits to get from one string to another by using the three above mentioned operations: substitution, deletion and insertion of a single character. For calculating the word error rate these operations are defined on single words instead of characters. Given the numbers of correct and erroneous words, the word error rate can now be calculated as follows:

$$WER = \frac{S+D+I}{S+D+C} \tag{2.51}$$

Due to insertions the numerator can be larger than the denominator. Thus, the *WER* is always equal or larger than zero and can also exceed one. Another measure directly related to *WER* is the word accuracy rate, which is defined as follows:

$$WAR = 1 - WER = \frac{C-I}{S+D+C} \tag{2.52}$$

As *WER* can be larger than 1 the word accuracy rate can become negative.

For an acoustic evaluation of a phoneme based recogniser, the phoneme error rate (*PER*) or the phoneme accuracy rate (*PAR*) can be used instead. It is basically calculated as mentioned in Equation 2.51 or 2.52 but considering each phoneme to be a word on its own.

Chapter 3

Speech Corpora

Appropriate data for training and evaluation of robust ASR systems and for understanding effects caused by acoustic distortion is crucial. Considering Figure 2.8 we can see that we can have different aspects influencing the speech signal from speaking characteristics towards channel characteristics. If we want to evaluate the effect of additive noise, we need data which offers additive noise characteristics but ideally avoids all other sources of distortion. However, if we want to evaluate channel effects, we should avoid additive noise and all other variabilities that might also influence the speech signal. In reality such ideal conditions for specific evaluations can hardly ever be achieved, so that we usually aim at minimising influences that are out of scope of our research.

One common way to provide data for evaluations of certain influences are simulated data sets. Such approaches are usually much less expensive than recording controlled realistic data. Furthermore, it is easy to avoid unwanted influences from other sources of distortion as only the desired aspects are simulated. Thus, we have perfectly controlled data sets with the acoustic distortion of interest in all desired levels and variations. The most commonly simulated type of distortion for research in robust ASR is additive noise and many approaches are only evaluated on such simulated data due to several advantages:

- Simulated noisy data with different levels of distortion are rather easy to collect. For additive noise, for example, clean speech data and noise samples can easily be mixed with different signal-to-noise ratios (SNRs).
- Considering additive noise, the SNR can exactly be determined for simulated data whereas the SNR can only be estimated for realistic noisy data, as clean speech and noise is not available separately.
- Clean speech and noisy speech of all SNRs are identical except for the added noise, so that a direct evaluation of the influence of only the additive noise is possible.

A major issue of such simulated evaluation sets, however, is that they are not necessarily as close to the realistic conditions as desired, which we will show later in this work. This is often caused by wrong assumptions in the simulation model or by data or parameters insufficient for realistic simulations. For additive noise, for example, the Lombard effect is present in realistic noisy data but difficult to simulate in artificial noisy data. Still, we will also consider two evaluation sets based on (partially) simulated data in our evaluations. The first set is the Aurora 2 evaluation set often used for evaluation of robustness approaches in ASR. The second is a step-by-step channel simulation of the Terrestrial Trunked Radio (TETRA) channel. TETRA is the digital radio communication standard for security forces used in many parts of Europe and Asia. Hardware, filtering, coding and transmission effects of this communication channel influence the speech signal and the speech recognition accuracy in various ways, which we will analyse and understand in experiments on this corpus later in this work.

Another way to get evaluation data for separate experiments on different sources of distortion is a well designed and carefully recorded speech database. One such example is our MoveOn Corpus. Due to a lack of corpora fulfilling the requirements of realistic distorted speech and separability of the sources of distortion as mentioned above, we designed, recorded, and finalised this corpus. It offers an evaluation set for analysing and understanding various aspects of acoustic distortion and their influence on ASR. Including synchronously recorded close-talk and throat microphone signals and realistic background noise from recordings on the motorcycle, this database includes relevant acoustic conditions for the planned analyses and evaluations. Thus, the MoveOn Speech Corpus will be the main corpus for our experiments in the following chapters.

While one of the reasons for selecting these three corpora is the difference of the provided acoustic distortions, another reason is the different level of complexity of the ASR task for which the corpora are designed. Aurora 2 is design for simple connected digit recognition with a vocabulary of 11 words, providing clean speech and speech with simulated additive noise. The speech and noise data in the MoveOn Corpus focuses on the more complex task of command and control with a vocabulary of more than 100 words using monophone acoustic models. The provided acoustic conditions include clean speech and speech with realistic background noise (partially with Lombard effect) as well as two rather different, synchronously recorded microphone channels. The TETRA Corpus enables the most complex task of large vocabulary continuous speech recognition with triphone acoustic models and a vocabulary of about 200,000 words. It offers clean speech and distorted speech for simulated and realistic distortions caused by the TETRA transmission channel.

In the following we will introduce the three evaluation sets in more detail. We start with the publicly available Aurora 2 dataset developed by Pearce and Hirsch, which enables the comparison with other approaches as it is widely used in the scientific community. Furthermore, we will detail the two sets that we specifically designed to enable experiments on certain aspects of acoustic mismatch and ASR. The first set is the simulated TETRA speech corpus based on clean speech TV broadcast with a step-by-step degradation caused by different influences of the TETRA radio transmission. The second set is the MoveOn Corpus. We will present the design and development of this corpus for the purpose of the MoveOn project and our research presented in this work. The creation of such a purposely designed corpus requires knowledge of and attention to various aspects from design towards annotation and is — as already mentioned above — much more effort than the simulation of certain influences and conditions. In the final part of this chapter we will compare the introduced corpora in terms of ASR complexity and acoustic distortion.

3.1 AURORA Project Database 2.0

The AURORA Project Database 2.0 (Aurora 2) is a publicly available[1] evaluation corpus for robust ASR on simulated noisy speech. It is used as a reference corpus in many scientific publications.

The Aurora 2 evaluation corpus ([47]) is designed for evaluation and comparison of the performance of algorithms for robust ASR. The corpus is based on TIDIGITS ([102]), a database of clean speech recordings of connected digits recorded at Texas Instruments, Inc., in 1982. The purpose of TIDIGITS was to support the development and evaluation of algorithms for ASR for connected digits. While TIDIGITS offers 77 recorded digit sequences from each of 326 speakers (111 male, 114 female, 50 boys, 51 girls) recorded in a quiet environment, Aurora 2 uses part of these training and test sets to add simulated additive noise and filter characteristics from the telecommunications domain. The data is provided at a sample rate of 8kHz. The language of the corpus is American English.

[1] Available at ELDA (AURORA/CD0002): http://www.elda.org/article52.html

	Noise type	Training set	Test set
N1	Suburban train	A,B,C	A,C*
N2	Crowd of people (babble)	A,B,C	A
N3	Car	A,B,C	A
N4	Exhibition hall	A,B,C	A
N5	Restaurant	—	B
N6	Street	—	B,C*
N7	Airport	—	B
N8	Train station	—	B
	** MIRS filter characteristics instead of G.712*		

Table 3.1: Noise domains of the Aurora 2 dataset. For evaluation on noisy speech, noise from different domains is added to clean speech utterances of connected digits in various SNRs. The defined training and test sets (A,B,C) of Aurora 2 contain noise from one out of the following eight noise domains (N1-N8).

3.1.1 Evaluation Sets

Noises for noisy speech simulation were recorded in eight different environments (see Table 3.1). Some of the noise environments mainly provide fairly stationary noise while others can contain a variety of non-stationary sound events as well. The different noise recordings are added to the clean speech TIDIGITS with different signal-to-noise ratios (SNRs) from 20 dB down to -5 dB in 5 dB steps. Furthermore, additional filtering based on the frequency characteristics of typical transmission channels in the telecommunications area (basically G.712 and MIRS; [103]) is applied. The SNR is calculated on the noise and speech signal filtered with the G.712 frequency characteristics beforehand.

Two training sets are defined as default: a clean speech training set and a multi-condition training set. Both training sets are based on the same 8440 utterances from 55 male and 55 female speakers from the TIDIGITS training set. In case of the clean speech training all utterances are filtered with the G.712 characteristics. For multi-condition training the 8,440 utterances are also filtered with the G.712 filter and divided into 20 subsets with 422 utterances in each subset. Each of the first four noise types (N1 to N4) in Table 3.1 is added to one of the 20 subsets in five different SNRs (clean, 20 dB, 15 dB, 10 dB, 5 dB).

Three test sets (A,B, and C) are created based on 4004 utterances from 52 male and 52 female speakers of the TIDIGITS test set. Four subsets with 1001 utterances in each of the subsets are defined.

Test set A adds one of the first four noises from Table 3.1 in 7 different SNRs to each of the subsets. The SNRs range from 20 dB to -5 dB in 5 dB steps and also include a clean speech version. Noise and speech are both filtered with G.712 characteristics before adding the signals. The noise conditions of test set A can be considered as “known” when using the acoustic models of the multi-condition training.

Test set B is created similar to test set A, but adds noises N5 to N8 from Table 3.1 instead of N1 to N4. Thus, test set B represents “unknown” noise conditions also for the multi-conditional acoustic models.

Test set C contains only two subsets with 1001 utterances. Instead of G.712, speech and noise is filtered with MIRS characteristics and noise from suburban train (N1) and street (N6) are added with the 7 different SNRs. Thus, the effect of acoustic channel mismatch in addition to additive noise can be evaluated.

Training (test)	Test set A	Test set B	Test set C
Clean (clean)	99.02%	99.02%	99.05 %
Clean (20dB-0dB)	61.34%	55.74%	66.14 %
Multi-conditional (20dB-0dB)	87.81%	86.27%	83.77%

Table 3.2: Aurora 2 baseline results as reported in [47]. Baseline experiments of ASR with clean speech acoustic models and multi-conditional acoustic models on the three evaluation sets A,B,C of Aurora 2 resulted in the presented word accuracy rates. The results are listed as the average on all clean speech test data without noise and all noisy speech test utterances between 0 and 20dB SNR as presented in the paper.

3.1.2 Reference Evaluation

A reference evaluation of ASR on the Aurora 2 evaluation set is also presented in [47]. The evaluation uses the Hidden Markov Model Toolkit (HTK) for training and recognition. The acoustic models are word models, i.e. each hidden Markov model (HMM) is modelling a full word (digits "one" to "nine" plus "zero" and "oh"). The HMM of each word has 16 modelled states and left-to-right topology without the possibility to skip a state. A mixture of three Gaussian distributions is used to model the probability distribution of each state's output. A diagonal matrix is assumed as covariance matrix. Additionally, a three state silence model and a one state short pause model with six Gaussian mixtures per state are included.

Typical MFCCs with 39 features are extracted using the first twelve cepstral values (without the zeroth coefficient), plus logarithmic energy, plus first and second order derivatives (Section 2.2.1). While in [47] also an alternative front-end standardised by ETSI is evaluated, we will focus on the non-ETSI evaluation here, which is similar to the baseline setup we will use later in this thesis.

Table 3.2 summarises the main results of the Aurora 2 baseline evaluation in terms of word accuracy rate. The performance measure used for each test set is defined as the average word accuracy rate over all noise types of each test set for SNRs between 20 dB and 0 dB. Unknown additive noise and different channel characteristics decrease the recognition performance in case of multi-conditional training. The word accuracy for MIRS filtered test data on clean G.712 filtered acoustic models seems to increase compared to test data without expected channel mismatch also filtered with G.712 characteristics. But as can be seen more clearly in the original paper, it is not possible to directly compare these averaged results of the different sets, as each set contains different types of noise with each type resulting in quite different word accuracies.

3.1.3 Summary

Aurora 2 offers an evaluation set for robust ASR for the rather simple task of full word digit recognition. This enables almost perfect word accuracy rates in clean conditions and still high rates for medium signal-to-noise ratios. It focuses on simulated additive noise only, but it is used in many scientific publications to evaluate and compare the performance of noise reduction and robustness algorithms for ASR. In our work we will use only Set A of Aurora 2 for reference evaluations as we focus on multi-conditional but matching conditions in Chapter 5.

3.2 TETRA Broadcast Corpus

Another source of distortion as discussed in Section 2.3 are channel characteristics mainly caused by hardware and transmission channels. One such channel is TETRA (Terrestrial Trunked Radio), a di-

gital radio standard by the European Telecommunications Standards Institute (ETSI) ([104]) design for robust speech transmission for public safety forces widely used in Europe and Asia. Due to a low bit rate of up to 7.2 kbit/s available for transmission, the TETRA codec is optimised for speech coding with high intelligibility at low bit rates. The TETRA implements an algebraic code excited linear prediction (ACELP) coding for speech, which enables speech transmissions down to 4.6 kbit/s. Thus, significant influences on the speech signal caused by the coding and decoding procedure can be expected. The TETRA transmission channel is interesting in this work for several reasons. First, we can expect significant influences on the speech signal and the speech features caused by several stages of an actual TETRA transmission from hardware effects and speech coding towards transmission effects. Second, we had access to actual TETRA hardware, which enabled us to record speech transmitted via a real TETRA channel, on the one hand, and to simulate the TETRA channel by using the standard software implementation of the TETRA codec, on the other hand. Thus, we can not only determine the influence of the entire transmission process on the speech signal, but we can also separate and estimate the magnitude of various aspects involved in the transmission in a step by step simulation.

As realistic TETRA communication data is hardly available, we decided to use recorded clean speech from another domain, which we transmitted via TETRA and re-recorded after transmission. This does not exactly cover the domain of public safety, but still enables an evaluation of the real transmission channel including a direct comparison with the original clean speech signal. We decided to consider the rather complex task of large vocabulary continuous speech recognition (LVCSR) here using a high quality German TV broadcast corpus as clean speech baseline. This way we are able to evaluate our approaches in Chapter 5 also on a complex task for ASR. Additionally, realistic TETRA communications can be considered to be rather unrestricted speech with a large but often domain-specific vocabulary, so that the complexity of the ASR task is close to the used LVCSR data.

Anyway, we are mainly interested in the acoustic effects of distortion in this work, so that we finally can neglect the specific lexical domain of the rescue operation. We use our large vocabulary TV broadcast corpus as a clean speech baseline to simulate the TETRA transmission channel step by step from frequency limitations towards transmission via a real TETRA channel. This extension of the clean speech ASR corpus enables us to evaluate the different sources of distortion of the TETRA channel on the LVCSR performance of our system as we will do in Section 4.4 and as partially published in [6, 7].

3.2.1 AM Baseline Corpus

The baseline for our simulated TETRA corpus is the Fraunhofer IAIS Audio Mining test and development speech corpus (AM Corpus). The corpus is designed for large vocabulary continuous speech recognition of TV broadcast shows. It currently contains almost 100 hours of fully transcribed German broadcast news and political talk-shows. Transcription of speech was performed and verified manually. With a focus on planned speech, it partially covers spontaneous speech as well. The original audio is sampled at 16 kHz and can be considered to be of high quality. Noisy sections of the recordings have been omitted and speech can be considered to be completely clean.

For evaluation purposes distinct training and test sets are defined. Table 3.3 shows the statistics for these training and test sets of the AM Corpus.

The size of this corpus enables a context dependent acoustic modelling using triphones commonly used to improve performance in LVCSR tasks. Instead of monophone models common triphone combinations are learned and are modelled by HMMs. Thus, the complexity of the acoustic models increases dramatically from one HMM for each of the about 50 monophones to one for each of the usually around 10,000 triphones. The aspect of increased complexity will be particularly interesting in Chapter 5 about blind acoustic model selection.

	Training set	Test set
Sentences	82 799	5 719
Words	723 933	46 978
Distinct words	52 700	8 799

Table 3.3: Statistics of training and test set of AM Corpus. The baseline corpus for creation of the TETRA Corpus was developed for training and evaluation of LVCSR systems. The corpus contains the listed number of sentences and words in the disjoint training and test sets.

3.2.2 TETRA-Extension of the Baseline Corpus

Even though the domain of the data (TV broadcast compared to TETRA communications) is rather different, acoustic effects can be analysed by extending the clean speech AM Corpus by simulated TETRA influences on the clean speech signal. The considered broadcast data is of high quality and without background noise. Furthermore, it allows high recognition rates for the complex task of LVCSR. So it is well suited to investigate the effects of acoustic degradation when used as baseline for a step by step simulation of the TETRA channel. In the following paragraphs we introduce the TETRA codec based on ACELP coding as standardised by ETSI. We further present the TETRA hardware available to us for recording real transmitted data. With this background information we detail the different levels of simulation and manipulation to create the different evaluation subsets of our TETRA Broadcast Corpus.

TETRA Standard

Terrestrial Trunked Radio (TETRA) is a standard, which was standardised by the European Telecommunications Standards Institute (ETSI) in [104]. It is designed for a digital and robust speech transmission for institutions of public safety.

The TETRA speech codec is based on the algebraic code-excited linear prediction (ACELP) coding model with special parametrisation. The algorithm of ACELP is closely related to CELP (code-excited linear prediction) using vector quantisation and a parametric coding. In general, the TETRA codec extracts and transmits on frame and sub-frame level linear prediction coefficients (LPC), algebraic codebook parameters with gains, and pitch. Therefore, for a given speech signal in 8 kHz, a pre-processing with offset compensation and downscaling of the speech signal is performed first. The pre-processed signal is then passed to the encoder. The encoder determines codebook and parameters by an analysis-by-synthesis technique on frames of 30 ms length. The LPCs are computed for each full frame. The coefficients are then converted to Line Spectral Pair (LSP) representation as detailed in the standard and quantised with 16 bit split Vector Quantisation (VQ). Further, algebraic codebook parameters as well as pitch and adaptive and fixed codebook gains are calculated for sub-frames of length 7.5 ms. A specific dynamic algebraic excitation codebook is used to concentrate the energies of the excitation vectors in the important frequency bands. The 16 bit algebraic codebook is determined by minimising the mean squared error between weighted input speech and weighted synthesised speech. The codebook gains are quantised by predictive Vector Quantisation. The combination of all parameters results in 137 bit/frame and consequently in a final bit rate of 4.567 kbit/s. On the decoder side, the received parameters are decoded and the decoded information is used to reconstruct the speech with a synthesis filter. For a complete overview, see [104].

Figure 3.1: Motorola CM 5000 radio station (left) and Motorola MTP 850 hand-held device (right). The TETRA equipment used to record actually transmitted speech is shown in these photographs. Clean speech from the AM Corpus is fed to the line input of the hand-held device, transmitted to the TETRA station, and recorded from the line output.

TETRA Hardware

For a real TETRA transmission we employed a TETRA radio station and a handheld device, which we used in direct mode to transmit speech from the station to the hand-held device and vice versa. The TETRA equipment available for our experiments is the Motorola CM 5000 radio station and the Motorola MTP 850 hand-held device (Figure 3.1). While certain hardware design decisions are dependent on the manufacturer, the general TETRA transmission with its encoding scheme complies to the ETSI standard described before. Thus, we can analyse the particular effects caused by this Motorola hardware, but in lack of other TETRA devices available for our experiments, we cannot generalise the specific results to all TETRA equipment. For our evaluations this is still sufficient, as we mainly want to analyse and identify effects on the speech signal that can be caused by hardware, but we do not want to create the detailed characteristics for all available TETRA equipment.

Simulation of TETRA Channel Effects

We create training and test sets for a step by step evaluation of the TETRA channel effects. The different evaluation sets are listed in Table 3.4. These sets were created as follows to cover several influences of a TETRA transmission:

Clean 16. This set is the original broadcast data with full 16 kHz sampling rate. It provides the baseline for ASR evaluations and presumably will provide the best results with lowest WER.

Clean 8. In this set we provide the original broadcast data down-sampled to 8 kHz. TETRA only considers speech data of 8 kHz sampling rate, which has certain effects on the ASR performance due to the low-pass filtering effect compared to the original 16 kHz signal. Recording and re-sampling of the audio signals was carried out using SoX.[2]

[2] `sox.sourceforge.net`

Name	Sampling rate	Characteristics
Clean 16	16 kHz	Original broadcast data
Clean 8	8 kHz	Downsampled broadcast data
AMR 4.75	8 kHz	*Clean 8* with AMR coding
TETRA Codec	8 kHz	*Clean 8* with software TETRA coding
TETRA Radio	8 kHz	*Clean 8* transmitted via TETRA hardware (line in)

Table 3.4: Overview on TETRA extension of the AM Corpus. Based on the AM Corpus (Clean 16) four different evaluation sets are simulated or recorded. A down-sampled version (Clean 8), a version with AMR 4.75 coding and a version with TETRA coding are provided. A transmitted and re-recorded signal is also available with TETRA Radio.

AMR 4.75. The clean speech, 8 kHz, data is further coded with AMR 4.75 coding. The Adaptive Multi-Rate (AMR)[3] speech codec has been used here since it features a comparable ACELP scheme as used in the TETRA coding (compare Section 3.2.2). AMR with 4.75 kbit/s is also mentioned as alternative codec to the TETRA codec in the ETSI standard,

TETRA Codec. This set contains clean speech 8 kHz data encoded and decoded with a software TETRA codec. For this purpose we use the TETRA codec reference implementation as provided by ETSI.[4]

TETRA Radio. In the TETRA radio set clean speech 8 kHz data is transmitted via a real TETRA channel using line input of the TETRA device. A notebook is connected to the line in of the Motorola MTP 850 TETRA hand-held device, transmitted to the Motorola CM 5000 TETRA station and recorded from the line output of the station on another notebook.

All sets use the same training and test set splits with the amounts pictured in Table 3.3 as exactly the same signals just modified in one way or another are used. Thus, simulation of the various steps enable an estimation of each step's influence on the ASR performance as we will show in Section 4.4.

3.2.3 Summary

Our extended AM Corpus enables evaluations on the influences of the transmission channel effects on ASR. With the step by step simulation, we are able to estimate, which aspect of the TETRA transmission degrades the speech recognition performance to which degree. Such an experiment will be presented in Section 4.4. Furthermore, the complex LVCSR task with triphone modelling enables an additional perspective on the presented approach of acoustic model selection in Chapter 5 in comparison with the less demanding ASR tasks of digit recognition and command and control with word and monophone HMMs.

3.3 The MoveOn Motorcycle Speech Corpus

The MoveOn Motorcycle Speech Corpus (or MoveOn Corpus) is a British English speech and noise corpus purposely designed for research of robust speech recognition on realistic noisy speech data. It

[3] www.3gpp.org/ftp/Specs/html-info/26104.htm
[4] pda.etsi.org

was created from scratch in the project MoveOn[5]. Due to the scope of the project, which included research in robust ASR, we were able to purposely design the corpus in all major aspects relevant in this work. The targeted ASR task is command and control with small to medium sized vocabulary. With the project background of developing a command and control and dialogue interaction system for police motorcyclists, speech data was recorded on the motorcycle to enable evaluations on realistic and domain-specific data. We will detail the design of the MoveOn database, which is the central evaluation corpus in this work.

As mentioned in Section 2.3 any acoustic mismatch between training and operational conditions shows in a decreasing speech recognition performance. As the sources of acoustic mismatch can be manifold, one of the goals of the database development was to avoid any unwanted mismatch, on the one hand, and to capture or influence particular sources of mismatch and distortion, on the other hand, to be able to evaluate several major sources separately. This includes, in particular, factors such as background noise and microphone channel effects. Furthermore, the objectives of the MoveOn project, which aimed at the development of a noise robust dialogue system operating on the motorcycle in a noisy outdoor environment ([105]), had to be taken into account and, thus will be detailed further down.

In [1] we outlined the main design of the database and offered a preliminary account for the collected speech and noise recordings. In the following, we will go further into details, introducing the requirements of the MoveOn project and related work on similar corpora to show the necessity for developing a new corpus. Afterwards, we offer a comprehensive description of the completed database with detailed information about statistics, annotation procedures and baseline results to show the versatility but also the limitations of this evaluation corpus. The full design and implementation description of the corpus was also submitted for publication in [2].

3.3.1 Project Requirements

Among the main objectives of the MoveOn project was the development of a robust multi-modal and multi-sensor low-distraction dialogue interaction system that supports information access and operational command-and-control protocols for the two-wheel police force in the UK ([105]). The information support is obtained either remotely from the control centre in the police station, or locally through the functionality provided by the wearable computing environment developed within the project. This environment offers several functionalities such as navigation support, accessing local information, storing video and audio streams for reporting and evidence collection purposes, automated logging and diary capabilities, information recall and storage on request, visualisation and alert mechanisms, communication with colleagues on the road or in flying vehicles, etc. The remote information access guarantees command, control, and guidance support as well as access to forensic and other police databases located at the central police station.

The multi-modal user interface developed for the MoveOn application consists of audio and haptic inputs, and audio, visual and vibration feedback to the user. Due to the specifics of the MoveOn application, involving hands-busy and eyes-busy motorcyclists, speech is the dominating interaction modality, especially when the user is on the move.

Despite the development of a noise-robust helmet, when driving in high speed, i.e., when patrolling the area on a motorcycle, chasing other vehicles, etc., especially engine and wind noise can severely affect the interpretation of the spoken commands. In these occasions the MoveOn system must be capable of delivering the required information in a non-obtrusive manner or — if this is not possible — switching back to human to human interaction as currently used by motorcycle police forces. The

[5] MoveOn (IST-2005-034753) is a Specific Targeted Research Project of the European Union's Sixth Framework Programme: Information Society Technologies (IST).

objectives can be fulfilled only through a careful design of the spoken interface and accounting for the noise conditions of the real-world environment. The latter required the development of a corresponding speech and noise database. These data served for creating the acoustic models of the speech recogniser, and for analysing, modelling and compensating effects of additive and other interferences. Summing up, the MoveOn Corpus has the purpose of providing representative speech and noise data, typical for the domain and environment, enabling a successful development and testing of acoustic modelling and noise reduction approaches for a reliable human machine interaction on a motorcycle.

3.3.2 Related Work

Similar requirements motivated the development of various speech databases which address the needs of different application domains. The databases often focus on a certain domain and certain noise conditions, but do not suffice our demand in terms of separated channel or background noise influences. Some of the more popular speech databases in a vehicle environment are described in the following paragraphs.

In [106] the design and development of a Japanese speech corpus recorded in a car environment is described. The purpose of the database was research and development in intelligent transportation systems, and the creation of robust dialogue and speech recognition systems in a noise environment. The linguistic content of the corpus was chosen with respect to phonetic balance of the resulting speech material. Recordings took place in a specially-equipped vehicle using the Wizard of Oz experiment simulating human-computer interaction. While the typical acoustic conditions were covered no particular considerations concerning evaluations of noise robustness and acoustic mismatch influenced the design decisions.

A Korean corpus for car environment with the purpose of research and development of improved ASR performance in such special conditions is described in [107]. Three types of cars and numerous environmental setups were used and speakers were chosen with respect to relevant characteristics for ASR including gender and accent of the speakers. The linguistic content focused on terms and phrases typical for the domain of in-car applications.

The authors in [108] set a focus on multi microphone and multi modal approaches for robust speech recognition in car environment. An array of eight omnidirectional microphones as well as four cameras captured speech with background noise and visual information from inside the car. The database consists of isolated digits and letters, phone numbers and phonetically balanced sentences.

One of the most popular and largest databases recorded in the vehicle environment is SpeechDat-Car ([109]). This database includes several different languages with terms and phrases for tasks such as voice-dialling, car accessories control, etc. Recordings were made in various conditions typical for the car environment and the design also put emphasis on a high degree of phonetic and speaker variability. Typical hardware for the relevant real-life applications was used for recording the database. So one of the main focuses was the development of a database capturing realistic data that is as close as possible to speech data seen in a final application.

In the particular domain of speech recognition on the motorcycle only little work has been conducted so far. One such effort was made within the SmartWeb project. The German SmartWeb Motorbike Corpus ([110]) was recorded under the special circumstances of a motorcycle ride. The idea of the SmartWeb system is that a motorcyclist can retrieve information related and dependent on his current activity. The recording setup tried to simulate as close as possible a real human-machine interaction with the SmartWeb system. The hardware used to capture the speech signal was chosen with respect to the limited available space and included a noise robust throat microphone for some of the recordings as well as a motorcycle helmet with communication equipment (microphone and headphones) connected to the

Item Code	Description	No. of items
AW001-AW065	Application words-phrases	65
BD001-BD005	Sequence of 5 isolated digits	5
PL001	Plate number	1
ID001-ID010	Single isolated digit	10
TP001	Time phrase	1
GW001-GW026	General words	26
LC001-LC014	Call signs	14
MW001-MW011	Special mandatory words	11
MS001-MS015	Special mandatory words-synonyms	15
OW001-OW022	Optional words-phrases	22
CP001-CP007	Confirmation phrases	7
SR001-SR010	Phonetically rich sentences	10
SP001-SP010	Spontaneous questions	10

Table 3.5: Content of MoveOn prompt sheets. The MoveOn Corpus contains various types of utterances related to routine operations of police units as well as general command and control utterances, phonetically rich sentences and spontaneous answers. The number of items per type and prompt sheet is listed in the table.

recording hardware via bluetooth. Additional distortion caused by the Bluetooth setup was reported.

While only the SmartWeb Motorbike Corpus fulfilled most of our requirements, the severe distortions due to the wireless Bluetooth communications and the differences in the linguistic design made the corpus insufficient for the MoveOn project and our evaluations.

3.3.3 Design

The corpus is designed to collect both (noisy) speech data as well as pure noise samples. The database covers a variety of different driving and environmental conditions from a realistic acoustic environment, recorded while professional police officers were performing simulated patrolling activities as well as additional clean speech recordings in an office environment. Our design of the linguistic content targeted at the coverage of application-specific commands, along with phonetic balance. The recording equipment was chosen with respect to the application design and the particularities of the 2-wheel vehicle driving environment.

Linguistic Content

The linguistic content of the database covers terminology and expressions related to the specific communication protocol used during a routine operation of the police force: command words and phrases, application words and phrases, and most of their synonyms. For guaranteeing sufficient representation of all phonemes, a number of phonetically rich sentences, taken from the British English SpeechDat(II)-FDB4000 database ([111]), were included in the prompt sequences.

Table 3.5 presents the structure of an exemplary MoveOn prompt sheet, where the items AW001-AW065 are application-specific words and phrases. In order to increase the frequency of appearance of the specific items, they appear twice in each prompt sheet. The items BD001-BD005 are randomized sequences of five isolated digits in one utterance. Ten spontaneous answers to questions stated in the prompt sheet are expected from the speaker (SP001-SP010). All the items described in Table 1, except the SR items (phonetically rich sentences), are common for all audio prompt sheets. The SR items

are different for each audio prompt sheet for guaranteeing sufficient frequency of occurrence even of rare phonemes. Further, all items are randomly distributed within each prompt sheet, ensuring a fair distribution of the prompts in respect of the changing noise conditions along the static predefined route used for the recording sessions.

Design of the Audio Prompts

The nature of the application setup with limited availability of manual and visual senses of the speaker while driving a motorcycle, brought up several challenges when coming to the implementation of the prompt-sheet. A previous attempt for creating databases in the vehicle environment ([106]) guided the subject by prompting phonetically rich sentences through a headset, while the whole procedure was controlled by the operator. A similar technique was followed in [110], where the person was instructed to imagine a certain situation connected with a task to solve. However, in the present case, the motorcyclist had to repeat the prompted word or phrase heard from the earphone attached in the helmet. For that purpose all the prompt items were recorded in studio environment by a native speaker of British English. In total, twenty-three prompt-sheets were created. Each prompt sheet starts with a short introduction, informing the speaker about the procedure he has to follow. Each prompt starts with a short phrase which introduces the motorcyclist if he has to repeat an utterance or answer spontaneously to a question. Every prompt ends with a DTMF tone, after which the speaker is expected to speak. The silence between two prompts lasts twice the total duration of the current prompt, ensuring that the speaker will have sufficient time to pronounce the utterance once the driving task permits. Each prompt sheet consists of 302 prompts, obtained by the duplication of the AW items plus 10 repetitions of the SP items plus the rest of the items. The resulting length of a typical prompt sheet is approximately 85 minutes.

Recording Equipment

In [110] various problems and drawbacks with the equipment were reported when recording a similar database for German speech on a motorcycle. Thus, we defined rather strict requirements for the database and the recording equipment to realise high quality speech and noise recordings avoiding unwanted distortions. We decided to use three microphones: two close-talk microphones attached inside the motorcycle helmet and a throat microphone placed around the neck of the speaker. While the close-talk microphones provide standard speech recordings with good frequency response, the throat microphone synchronously captures the same speech signal with less environmental noise but different microphone characteristics.

In order to cope with the limited space in the helmet, the close-talk microphones needed to be small and lightweight. Furthermore, a good and almost linear frequency response in the relevant spectrum of speech was desirable. Directional microphones and specific frequency responses, which are often used for microphones in adverse environments, were not considered in order to avoid effects on the natural speech and noise spectrum. The close-talk microphones must further provide low distortion for high acoustic pressure levels in order to achieve speech signals of sufficient quality even under the extreme noise conditions anticipated on a motorcycle. We decided to use the miniature lavalier microphone AKG C417, which successfully fulfils these requirements. It has omnidirectional characteristics and does not require a certain direction towards the mouth. This avoids problems with wrong adjustment of the microphone, but does not reduce environmental noise when recording speech. While this is usually a disadvantage in ASR, we prefer such characteristics in our case as it enables us to collect undistorted environmental noise as well. However, the lack of noise reduction does not seem to influence the

intelligibility too much, which can be contributed to the near field speech capturing in combination with the closed acoustics of full-faced helmets used in the recording campaigns.

An additional throat microphone provides a speech signal, which is almost free of any additive interference from the environment at the cost of a band-limitation of the speech signal. The throat microphone is placed at the throat directly picking up vibrations produced by the larynx instead of capturing air-borne sound. Major requirements for the throat microphone are robustness towards mechanical stress but sensitivity for vibrations caused by the speech articulators. Unfortunately, at the time of the corpus design specification process hardly any throat microphone with detailed specifications was available. Thus, we based our decision on a preliminary test of two available models, the Tork Max Throat Microphone and the Alan AE 38 Throat microphone. In brief, the Alan AE 38 is a single transducer throat microphone with a neck strap to fixate the transducer on the larynx, the Tork Max microphone offers a similar construction but a dual transducer concept. Both microphones were only available with a proprietary connector, so that we build an adaptor to connect the microphones to the standard microphone input of our recording device. Due to undefinable blackouts of the Tork Max Throat Microphone, we decided to use the Alan AE 38 Throat Microphone.

The recording device must be small and lightweight but suited for high quality recordings with a minimum of three microphone inputs and recording channels. We decided to use the ZOOM H4 recorder, which offer an appropriate size, good technical specifications and flexibility of the microphone inputs including support of the bias required for the AKG microphones. Data storage and battery power are sufficient for a full recording session of up to 90 minutes. Unfortunately, the ZOOM H4 only supports two microphone channels, so that we used two devices in parallel.

Equipment Setup

In Figure 3.2 we summarise the recording setup. Both helmet microphones were connected to the first recording device. In addition, a throat microphone was placed around the neck and positioned on the throat of the motorcyclist, so as to capture the vibrations from the larynx. A wrong or loose adjustment of the transducer of the throat microphone can lead to a distorted signal and must be avoided. This is a problem especially for very small necks, as in this case the neck strap does not provide enough pressure to make sufficient contact between transducer and larynx. In our setup the throat microphone was connected to the second audio recording device. The ear phones delivered with the throat microphone were used to play back the audio prompt and were connected to the output of the first device (audio prompt channel). In addition to the three microphone channels, a channel superimposing the audio from the pre-recorded prompts and the audio from the close-talk microphones was recorded for the needs of a precise synchronization and in support of the annotation process. As in [110] distortion caused by the Bluetooth connection between helmet microphones and recording device was reported, we used a wired setup of the equipment to avoid this source of distortion.

The hardware setup was preliminarily tested in laboratory environment and on a small motorcycle in a realistic environment. During this process the interaction of all devices as well as the mechanical resistance of the single components and the entire setup were tested and improved. Cabling and setup were also evaluated in terms of safety and driver's distraction to avoid additional endangerment of the motorcyclist. Potential pitfalls and problems during the preliminary tests were noted in a check list to avoid these problems during the recording campaign.

The recording level of both devices was adjusted carefully to avoid clipping of the speech signal. We arranged the close-talk microphones about 4 cm left and right from the mouth of the speaker in the motorcycle helmet. The distance slightly varies from helmet to helmet. Both close-talk microphones are fixated with hook-and-loop tape on the cushion of the helmet. Hence, the equipment can easily and

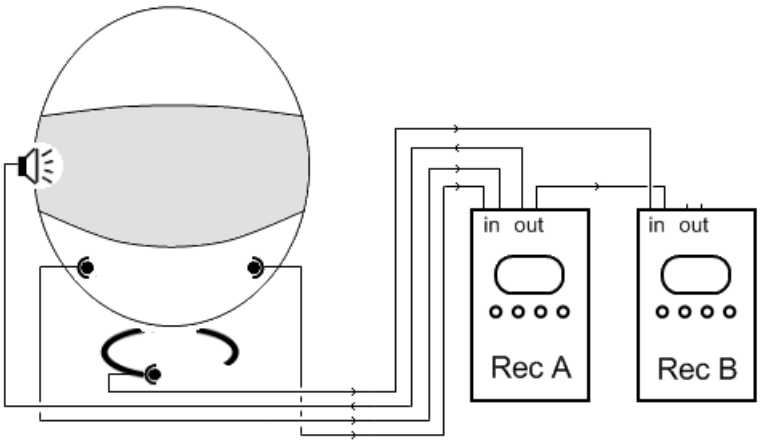


Figure 3.2: Hardware setup for the MoveOn database recording sessions. The close-talk microphones are installed in the helmet and the throat microphone is placed around the neck. The two ZOOM H4 recording devices record the close-talk microphone signals (Rec A) and the throat microphone signal as well as a synchronisation channel (Rec B). The prompts are played back from device Rec A.

reliably be adapted to the different helmets used during the recording campaigns. The throat microphone was put around the neck of the motorcyclist with the provided neck strap. The recording devices, as well as supplementary equipment and cables, were stored in a backpack to keep the setup independent of the motorcycle and for guaranteeing the safety of the motorcyclist.

Although a sample rate of 16 kHz is sufficient for ASR, we recorded the entire database at a sample rate of 44.1 kHz with 16 bits resolution. This can be particularly interesting for capturing the properties of the noise environment to enable research of the adverse noise conditions.

Documentation and Briefing

We prepared several forms and questionnaires based on recommendations in [112], including a speaker protocol and a session protocol. The speaker protocol covers all relevant information about each speaker, such as information related to pronunciation accent, age, gender etc., and the session protocol covers information about the particular recording session. An introduction describing the idea and background of the data collection campaign was offered to each speaker to improve the comprehension about the task and the way to act and speak during the recording session. A recording manual and check lists were prepared to support the supervisor before and during the recording sessions. All necessary forms were handed to the supervisor a week in advance in order to familiarise with the procedures. During the recording campaign all speakers were introduced to the recording procedure and — in case of the motorcycle recordings — to the route by the session supervisor. After completing each session all required information about session and speaker was filled into the protocol forms. Table 3.6 offers details about the nature of these information.

3.3.4 Implementation

We prepared and accomplished audio recordings in two different environments: on the motorcycle in a realistic environment and in a clean office environment for clean speech reference recordings. All data was annotated including spoken content and background noise.

Speaker Protocol	
Basic Information	Date of birth, sex, handedness, height, weight
Mother Tongue	language, dialect, place of elementary school, language of mother, language of father
Experiences as Motorcyclist	Years of riding a motorcycle
Session specific	Session ID, glasses, smoker, piercings, props, beard
Comments	*free for comments*
Session Protocol	
Basic Information	Session file name, speaker ID, prompt ID, date and time, hardware setup, supervisor
Scenery	Type of prompting, motorcycle, helmet, weather conditions
Route	Route, familiar with route, deviations, breaks, prompts finished
Noise and Environment	No. of missed prompts, traffic level and background noise (entire rout), cellular phone off, technical problems
Comments	*free for comments*
Traffic Conditions and Deviations from Route	Classification of traffic conditions, comments, deviations (for each section of the route)
Noise Conditions and Technical Problems	Classification of noise conditions, comments, technical problems (for each section of the route)

Table 3.6: Content of speaker and session protocols. For each recording session general speaker and session information was filled into two protocol forms. The covered information includes recommendations from [112] as well as additional MoveOn specific feedback on setup, background noise and recording route.

Recording Campaigns

In a first campaign, speech was recorded on a motorcycle in a realistic environment. Several types of motorcycles and helmets were used during the data recording campaign, most of them typical for the British police forces. The list of motorcycles includes amongst others BMW RS1200, Honda Pan European, BMW K1100, and Honda GoldWing GL1800. Furthermore, we experimented in different sessions with a variety of helmets (e.g. Shoei XR1000, Schuberth C2, Shoei Multitec, etc.), which cover the typical helmets used in the daily routine.

We defined the protocol of the first recording campaign with special care to capture the operational environment and the domain of the MoveOn application. For that purpose we chose a controlled environment in terms of a fixed route through the city and suburbs of Birmingham, UK. The route contains major environmental conditions, e.g., major and minor city roads, motorways, tunnels and country roads. Such a route enabled a more convenient assessment and interpretation of the various environmental noise types and ensured a sufficient coverage of the major noise conditions. The route was used in both directions and the sequence of the prompt items within the prompt sheets was shuffled in order to guarantee that in different sessions a specific utterance is recorded in different environmental conditions. A video of the route was recorded in support of the database development and as a review of the characteristics of the chosen route.

For the needs of the first data collection campaign, professional police motorcyclists from West Midlands Police, UK, were recruited. All speakers were native speakers — with and without area-specific pronunciation accent. Recruiting experienced motorcyclists was considered important for guaranteeing safety of the speakers, since their routine makes them less susceptible for mistakes caused by additional distraction and workload. Furthermore, selecting experienced police officers enabled a broader understanding of police procedures and protocols, as well as common terms of communication between police officers, on which the database is mainly based. Both qualifications contributed to the quality of the database. However, a disadvantage which came in consequence of our decision was the fact that hardly any female police motorcyclist can be found, so that only male speakers were available for the outdoor recording campaign. A total of 40 recording sessions with 29 different speakers were accomplished on the motorcycle.

In a second campaign additional sessions were recorded in a silent office environment using an identical hardware setup including the motorcycle helmet. The purpose of the second campaign was a reference collection of clean speech data.For the needs of the second data collection campaign we recruited additional British English speakers. Six male and four female native speakers — partly with an area specific pronunciation accent — were selected. Except for the speakers and the environmental setup, i.e., choosing an office environment instead of the realistic motorcycle environment, we kept all parameters of the second campaign identical to the parameters of the outdoor campaign. Thus, we recorded 10 additional sessions with 10 different speakers in an office environment.

Annotations and Validation

The annotation of the MoveOn noise and speech database was realised in two parallel procedures: annotation of speech and annotation of background noise. The annotation process was performed using Praat ([113]). For each session, two different annotation file-templates were given to the annotators — one for the noise annotation and another for the speech annotation. In Table 3.7 we show a schematic overview of the annotation structure with reference to the annotation tiers used as well as some details about the information expected to be filled by the annotators. More details about each of the two annotation processes are described in the following sections.

Speech	
Words	*Word-level transcription*
Affect	Positive-active (posa), positive-passive (posp), negative-active (nega), negative-passive (negp), neutral (neu)
Speaker	*Visual help for annotators showing segments where to expect speech*
Noise	
Air wind noise	low (a+), medium (a++), high (a+++)
Engine noise	low (e+), medium (e++), high (e+++)
Other noise	traffic, rain, tunnels, ...
Sound event	Horn, (passing) vehicle, ...
Visor	open, closed
Speaker	*Visual help for annotators showing segments where to expect speech*

Table 3.7: Annotation structure of the MoveOn Corpus. Both speech and noise of the recorded sessions were annotated using the tool Praat. Speech annotations included an orthographic transcription and the speaker's affect. Noise annotations considered dedicated tiers for the dominant sources of noise air wind and engine, and general annotation tiers for sound events and (continuous types of) other noise. Visor enabled the annotation of the position of the helmet's visor.

Speech Annotations Three tiers are used for speech annotation: a Speaker tier with automatically extracted prompt boundaries for visual support of the annotators as well as a Words and an Affect tier.

Research in speech recognition points out that speech recognition performance is also affected by the underlying affect in speech ([60]). Thus, the inclusion of the tier Affect was considered during the design of the speech annotation. The annotators were asked to define the area in the activation-evaluation space ([114]) where the affective state of the motorcyclist can be placed based on their human intuition: Positive-Active (posa), Positive-Pasive (posp), Negative-Active (nega), Negative-Passive (negp) and Neutral (neu). The first dimension of the label describes *valence*, the degree of pleasant or unpleasantness (here positive, negative). The second dimension is *arousal*, which is the degree of activity caused by an emotion (active or passive). Angriness, for example, is Negative-Active. The annotation of the tier Affect revealed only a small number of utterances with emotional data (posa (39), posp (52), nega (9), negp (52)). All the remaining instances were marked as neutral (neu). This low amount of non-neutral utterances can be explained by the fact that the speakers were not asked to act, as our main objective was to collect naturally occurring emotional speech, as it occurs during the patrolling activities.

In the tier Words the exact boundaries of speech were marked and the transcriptions of the spoken content were added by the annotators. During the annotation process we followed the SpeechDat conventions ([111]) for denoting word truncations, non-understandable speech and non-speech acoustic events. The lexicon of the speech database was created inheriting British English SpeechDat conventions with SAMPA phoneme transcriptions ([111, 115] — also see Table A.1). In Table 3.8 we present the phoneme frequencies in the speech corpus. The following phonemes were rare: OI, U@ and Z. Overall, the database design specifications about minimum frequency of appearance of each phoneme were achieved.

In Table 3.9, the number of items per category and their respective duration in seconds are presented. Here, RU items correspond to the out-of-prompt sheet transcriptions or to transcriptions where it was

Phoneme	Total Number	Phoneme	Total Number
@	8017	r	6146
D	1142	s	9503
I	10886	t	12805
N	1190	v	3494
Q	3825	w	1711
S	1320	z	2381
T	1229	{	3365
U	778	3:	852
V	1741	@U	5834
Z	156	A:	1132
b	1821	I@	1000
d	5703	O:	1430
e	5037	OI	36
f	3897	U@	10
g	1370	aI	3714
h	1581	aU	810
j	1046	dZ	1393
k	7910	e@	169
l	7921	eI	5503
m	4937	i:	3273
n	9322	tS	1373
p	5296	r	6146

Table 3.8: Number of phonemes in MoveOn Corpus. The final corpus contains all phonemes from the British English SAMPA phoneme table (Table A.1). The total number of occurrences for each phoneme is listed here. The phonemes U@, OI and Z are sparsely represented in the corpus.

Category	Total Number	Duration in Seconds
Application words-phrases	3587	7246
Sequence of 5 isolated digits	129	419
Plate number	28	110
Single isolated digit	291	199
Time phrase	29	92
General words	740	574
Call signs	390	804
Special mandatory words	327	530
Special mandatory words-synonyms	419	554
Optional words-phrases	600	1044
Confirmation phrases	190	241
Phonetically rich sentences	268	807
Spontaneous questions	1275	3647
RU items	2610	4708
Total number/duration of items	10883	20977

Table 3.9: Number and duration of recorded items per category in MoveOn Corpus. The items of the prompt sheets as introduced in Table 3.5 are represented with the listed amount in the final corpus. Number and duration were extracted automatically. RU items include all "random utterances" which could not be classified reliably due to a missing synchronisation channel.

not possible to identify the code. The latter occurred in sessions where the prompt channel, which is used to synchronise the sequence of the audio prompts with one of the in-helmet channels, was not recorded properly. In total, the completed database consists of approximately 6 hours of transcribed speech segments.

Noise Annotation We used six distinct tiers for the noise annotations: Air Wind Noise and Engine Noise for the respective noise types, Sound Event for temporary events (horn, passing car, etc.), and Other Noise for all other, rather stationary noises (traffic, rain, etc.). Furthermore, the state of the visor (open, closed) — if it was possible to determine — was marked in the tier Visor. The automatically generated tier Speaker (identical to the one in the speech annotations) provided visual support about the expected speech borders to the annotators.

A preliminary inspection of the data revealed a range of common noise types and sound events. Thus, for the dominant and most frequent noises, Air Wind and Engine Noise, a distinct tier was assigned. The annotators were asked to define the boundaries for the segments which contain such events, and assign intensity levels with one, two or three "+" symbols, according to their amplitude. All other noises were marked in the tiers Sound Event and Other Noise dependent on their characteristics. After defining the boundaries of the segments, each segment is labeled with one or several predefined noise labels from the preliminary inspection. As the state of the helmet's visor (open, closed) changes the acoustics, which was often clearly audible to the annotators, we introduced the additional tier Visor. The noise annotations were performed by different annotators but validated by a single person to achieve a consistent annotation for all sessions.

Table 3.10 shows the percentage of occurrence of the different noise types in the MoveOn database. The dominant background interferences were Air Wind Noise and Engine Noise, which usually coincide. The intensity of both types of noise is correlated as the intensity of air wind noise usually increases with increasing velocity of the motorcycle, i.e., with the engagement of the engine. All other types of

Air Wind	w/o	a+	a++	a+++	—
	48%	28%	15%	9%	—
Engine	w/o	e+	e++	e+++	—
	9%	63%	22%	6%	—
Sound Event	w/o	vehicle	noise	visor	others
	94%	2%	2%	1%	1%
Others	w/o	traffic	tunnel	noise	others
	83%	13%	2%	1%	1%

Table 3.10: Noise statistics of the MoveOn Corpus. The statistics show the occurrence of the most common noise categories in percent of recording time for the major noise annotation tiers. *w/o* states that no noise was marked in this tier. Major sources of noise include passing *vehicle*, unidentified general *noise*, noise when changing state of *visor*, and all remaining noise (*others*).

noise (annotated either in the tier Other Noise or Sound event) occur less frequently in the database, and their intensity is usually lower, when compared to air wind and engine noise. For instance, for about 83% of the recording time no Other Noise was reported as noise background and for about 94% of the recording time no Sound Event was reported, whereas air wind noise is present for more than 50%, and engine noise for more than 90% of the recording time.

Database Validation The validation procedure followed the existing standards ([111]) with adaptations to the needs of the MoveOn project as described below. A pre-validation procedure took place right after the first recordings and annotations. The result of the pre-validation process imposed repetition of the annotation of the first three sessions, towards correcting a variety of deviations from the predefined annotation conventions. Corrections and recommendations were communicated to the annotators to avoid further deviations in the annotation process.

Once the database annotation was completed and the annotations were checked by an expert, we performed database validation.Validation of the noise annotation included a semi-automatic check for consistent naming of the same noise types for all sessions and a manual check for correct annotations based on random samples of each recording session. Minor mistakes were corrected immediately. No major mistakes were uncovered in the noise annotation.

Initially, automatic error spotting of the transcribed items was performed for the speech annotation utilising the SpeechDat British English dictionary ([116]) enriched with the MoveOn-specific vocabulary. The significant number of mistakes found imposed a detailed inspection of approximately 10% of the collected data. This inspection uncovered certain types of common transcription errors in all sessions. Thus, the annotators where instructed to reprocess the speech annotations of the database on its whole accordingly. Revalidation of the later outcome was performed automatically following the same procedure utilised in the validation process. Results indicated absence of any mistakes in the speech transcriptions, thus the database was declared ready for organisation and distribution.

3.3.5 Organisation

The MoveOn database offers predefined training and test sessions after dividing the indoor and outdoor recordings into balanced subsets. We used a ratio of 80% training and 20% testing data. Specifically, given that for a limited number of sessions during the recording campaign data loss (in one or more audio channels) occurred, we consider the data completeness as an important criterion for performing a fair split of the database. Furthermore, since both the motorcycle and the helmet type affect the en-

		Full sets		Core sets	
		training	**test**	**training**	**test**
Main sets	right mic.	7995	1895	5533	1260
	left mic.	7572	1895		
	throat mic.	6530	1459		
Subsets	office	—	—	1535	397
	motorcycle	—	—	3998	863
	c&c	—	—	5533	465

Table 3.11: Default evaluation sets of the MoveOn Corpus. Several standard evaluation sets are defined for the MoveOn Corpus with the mentioned number of utterances per set. Main sets are the three different microphone channels. Subsets define parts of these sets with specific acoustic (office or motorcycle recording session) or linguistic conditions (command and control phrases only — c&c). Core sets contain only speech recordings available for all three channels.

vironmental conditions, they were considered to be important criteria for splitting the datasets. Finally, sex information was included as criterion, while age statistics were not considered due to the limited number of speakers in the database.

Based on the general database organisation described above, we defined various training and test sets and subsets to enable the evaluation of different aspects of robust speech recognition. The results were two major training and test sets per channel: first a complete set (full set) with all available data per channel, and second a core set containing only utterances that were recorded on all three channels synchronously. Thus, the core set is the most general test and training set enabling both a direct comparison of the performance of all three channels and an evaluation of robust speech recognition approaches making use of more than one microphone channel. Furthermore, several subsets can be defined, e.g., an office subset, a motorcycle subset or a command and control test subset containing only office recordings, motorcycle recordings or a test set with only command and control phrases (i.e. AW and CP items from Table 3.5 respectively. Table 3.11 shows the number of utterances for each evaluation set and subset. The number of utterances for each channel is lower than the total number in Table 3.9, as not all sessions provide all recording channels due to failures of the recording equipment.

3.3.6 Baseline Experiments

In several baseline experiments we evaluate the acoustic performance of an ASR system trained and tested on the sets and subsets of Table 3.11.

For evaluation a statistical approach for ASR on a phoneme level with hidden Markov models (HMMs) is used based on the Hidden Markov Model Toolkit (HTK). We decide for a setup widely used in ASR. Our acoustic models are trained from 39-dimensional feature vectors containing the first 12 static MFCCs (without the 0-coefficient) calculated for frames of 25 ms window length and a stepsize of 10 ms, plus energy and their first and second order derivatives. Cepstral mean normalisation is performed. Such feature vectors are widely used in the scientific community and for applications. In most of our evaluations in this thesis we will use this "standard" set of features. Each state of the HMMs is described by 16 Gaussian mixtures. A set of acoustic models contains a monophone model for each SAMPA phoneme (plus silence and short pause models). For each channel a separate set of acoustic models was trained using the channel specific dataset.

The acoustic performance is evaluated on a phoneme basis without using any lexical knowledge. Each phoneme of the SAMPA phoneme set with 44 phonemes was considered to have the same probability

training/test	full/full	full/core	core/core
Left channel	52.7	52.7	52.7
Right channel	52.0	52.6	53.0
Throat channel	46.8	45.8	45.8

Table 3.12: Comparison of phoneme accuracies for full and core evaluation sets. The main sets from Table 3.11 are compared by their average phoneme accuracies in percent. Several training and test set combinations for each of the three recording channels are evaluated.

of occurrence. Phoneme interdependencies and lexical knowledge were not considered to evaluate only the acoustic performance of the speech recognition system.

Full Set vs. Core Set

The acoustic performance of the speech recognition system was determined for the full and the core evaluation set in Evaluation I to determine the effect of the amount of data on the recognition accuracy. The core evaluation set has a reduced amount of data compared to the full evaluation set, as only sessions available for all three channels are considered. Thus, we first compare the results for the full and the core evaluation set for each channel to investigate the differences in the recognition performance, before we compare all three channels based on the core evaluation set.

In Table 3.12 the average phoneme recognition accuracies are presented for each channel and three different training and test set combinations. The first column shows the performance for acoustic models trained on the full training set and tested on the full test set for each channel. The second column presents the results for the same training set but tested on the core test set of each channel. The results for the core test set tested on acoustic models trained on the core training set are shown in the last column.

The phoneme recognition accuracy for the right and the left microphone channels is nearly identical, but the recognition performance on the throat microphone channel is distinctively lower. This matches our expectations, as left and right microphone are of the same type of high quality close-talk microphones, but the throat microphone provides a reduced signal quality due to the alternative transducer concept. We will see later that the reduced influence of background noise leads to a similar performance of the throat microphone compared to the close-talk microphone in case of motorcycle recordings only. The results for the different training and test set combinations for each channel are almost equal, especially comparing the last two columns with results based on the same test set. The higher amount of training data in the full training set compared to the core training set seems to have no major effect on the recognition performance. Hence, we will use the core evaluation sets for the following evaluations.

Office vs. Motorcycle Subset

We now test the office and motorcycle subsets of Table 3.11. This enables us to evaluate the influence of the environmental conditions on the recognition performance. Both subsets are recorded with the same hardware setup, but the office subset contains no background noise at all while the motorcycle recordings are from a realistic environment with a variety of background noises and noise levels. These subsets are rather small as can be seen in the table and might not be sufficient to train representative acoustic models. However, several effects based on acoustic mismatch can still be shown by this approach.

In Table 3.13 we present the results for each channel and subset on a speaker level. The acoustic models trained on the office subset perform best for the test speakers of the office environment (sessions 005 and 010) while the acoustic models trained on the motorcycle subset perform best on the test speakers

Session	Office subset			Motorcycle subset			Core set		
ID	**Left**	**Right**	**Throat**	**Left**	**Right**	**Throat**	**Left**	**Right**	**Throat**
005	58.6	57.4	37.5	47.4	46.2	34.4	57.3	58.5	38.0
010	56.5	52.5	35.0	28.7	25.6	20.5	49.9	47.9	32.0
107	25.3	22.6	37.3	56.8	59.5	61.6	53.8	55.9	57.9
118	33.2	32.9	40.0	59.4	61.6	53.8	57.6	60.2	52.1
126	28.8	23.5	34.0	54.3	52.7	49.1	51.6	50.3	47.2
139	28.1	25.7	41.1	46.4	46.4	46.4	46.1	45.1	47.3
mean	38.4	35.8	37.5	48.9	48.7	44.3	52.7	53.0	45.8

Table 3.13: Phoneme accuracies for each test speaker of the core evaluation set. The two test speakers from office environment (Sessions 005 and 010) as well as the test speakers from motorcycle environment (Sessions 107, 118, 126 and 139) are evaluated using the full core evaluation sets and the office and motorcycle subsets. The phoneme accuracies in percent are presented.

from the motorcycle domain (all other sessions). This effect is not surprising as there is no environmental mismatch between training and test set for these setups. However, the acoustic models for the throat microphone channel trained on the office subset show a similar performance on both office and motorcycle test speakers as opposed to the results for the close-talk channels. This can be explained by a rather small acoustic mismatch between both subsets due to the throat microphone technology, which does not capture as much environmental noise as standard close talk microphones. On the other hand, acoustic models trained on the motorcycle subset of the throat microphone data show an increasing performance for the speakers of the same environment but a lower performance for the speakers from the office environment. Thus, the environment still influences directly or indirectly the signal captured by the throat microphone. One of the influencing factors could be the style of speaking, which is usually influenced by the environment and environmental noise, for example, introduced by the Lombard effect ([64, 65]). Furthermore, the signal quality of the throat microphone depends on a proper adjustment of the microphone. A poor adjustment without sufficient constant pressure between sensor and larynx (especially for smaller necks of women such as in the case of the test speaker in session 010) reduces the signal quality considerably affecting acoustic model quality and recognition performance.

Command and Control Utterances

Next, we compared the phoneme accuracy rates for the core test set and the core test subset with command and control phrases only (please refer to Table 3.14) to evaluate, if the larger number of command and control phrases in the training set influences the recognition performance for the same type of utterances. We do not use any lexical knowledge here but only evaluate the acoustic performance. Compared to the core set with all utterances the command and control subset has about 10% absolute higher phoneme recognition rates for all channels. This can be explained by the design of the MoveOn database, which focused on command and control applications (AW and CP items) and thus provides higher phoneme frequencies for words typically used in these items. Consequently, phonemes that appear in the command and control utterances are better represented, and hence, their acoustic variabilities are usually better modelled than the ones of other phonemes like rare phonemes, which only occur in phonetically rich or spontaneous sentences of the database. More detailed investigations beyond an acoustic evaluation on the ASR performance for command and control also including lexical knowledge will be presented in Section 4.1.

The previous experiments for robust ASR present the covered aspects of distortion of the MoveOn

Speaker	Left	Right	Throat
Core set	52.7	53.0	45.8
c&c core set	58.7	59.1	50.5

Table 3.14: Comparison of phoneme accuracies of core set and command and control (c&c) test subset. The phoneme accuracy rates for the command and control test subset are compared to phoneme accuracy rates of the full core test set. Recognition on well represented phonemes of the command and control phrases show higher accuracy rates than evaluation on the full core test set.

Corpus and their influence on the recognition performance. The experiments further show the capability of the MoveOn database to serve as an evaluation test-bed with realistic data and provide a baseline for evaluations on robust ASR.

3.3.7 Summary

The design and implementation of the unique MoveOn motorcycle speech and noise database was focused on the needs of research in the field of robust ASR defined by the requirements of our work and the MoveOn project. However, the design of the database was kept sufficiently general which allows its use for a wider range of applications in motorcycle on the move environments. The speech and noise statistics show good coverage in terms of phoneme distribution and provide information about predominant types of background noise. The usefulness of the MoveOn database was illustrated in several exemplary evaluations presented in Section 3.3.6 focussing on difficulties of the particular environment and the different microphone channels. The database is in process of being released by ELDA[6] in 2012.

3.4 Comparison of Evaluation Corpora

We consider three different evaluation corpora with rather different levels of speech recognition complexity and acoustic distortion. The main characteristics of the three corpora relevant in this work are compared in Table 3.15.

The first corpus is Aurora 2, which is a well known evaluation corpus for digit recognition in (artificial) additive noise. Aurora 2 is an evaluation corpus for noisy digit recognition based on simulated additive noise and channel characteristics. In brief, it models full word HMMs for the low complexity task of digit recognition. The clean speech utterances are American English taken from the TIDIGITS database and changed according to the filter characteristics of two typical channels in the field of telecommunications (G.712 and MIRS). Noisy speech is created by adding noise samples from different domains to filtered clean speech in different SNRs. The set contains 110 different speakers. Its main advantages are the controlled noise conditions with different SNRs and a good comparison of the results when evaluating with this corpus as it is widely used for evaluations in robust ASR. The main disadvantageous are the small size of the corps, the very limited task and domain of digit recognition and the artificial character of the noisy speech.

The MoveOn Corpus with different realistic channel distortion and noise characteristics is used as another corpus for analysing effects of distortion and to evaluate the performance of ASR. The purpose of this corpus is the task of command and control, which has an increased low-to-medium complexity compared to Aurora's digit recognition task. A vocabulary of about 140 words, a basic grammar and a monophone based recognition introduces a higher degree of freedom and flexibility of the speech input.

[6] ELDA - Evaluations and Language resources Distribution Agency, http://www.elda.org/

Corpus	ASR Task (complexity)	Language (Task vocab.)	Modeled units	Corpus size	Sources of distortion
Aurora	digit (low)	American English (11 words)	words (11 + sil)	small (a few hours)	noise (simulated)
MoveOn	command & control (low/medium)	British English (140 words)	monophones (44 + sil)	small (about 6 h)	noise, microphones (realistic)
TETRA	LVCSR (high)	German (>200,000 words)	triphones (>5000)	large (>60 h)	channel (simulated, realistic)

Table 3.15: Comparison of evaluation corpora. The three evaluation corpora Aurora 2, MoveOn, and TETRA are compared by their main characteristics. Differences of the corpora and challenges for ASR are mainly caused by task complexity, vocabulary size, complexity of the acoustic models, corpus size and included sources of distortion.

The MoveOn corpus is a small corpus of realistic data. Three major sources of distortion are generally present in the corpus: speech and speaker variabilities (including Lombard speech), background noise and microphone channel distortions. For separate evaluation in the following chapter the latter two are of particular interest. Generally, we can divide the recorded data into four sets each one influenced by the presented sources of distortion in a different way:

1. **close-talk microphone, office:** clean conditions, good frequency response, almost ideal recordings, synchronously recorded with 2
2. **throat microphone, office:** clean conditions, distortion due to limited throat microphone capabilities, synchronously recorded with 1
3. **close-talk microphone, motorcycle:** clean conditions + additive background noise, partially with Lombard effect, synchronously recorded with 4
4. **throat microphone, motorcycle:** clean conditions + low amount of additive noise (due to throat microphone technology), distortion due to limited throat microphone capabilities, partially with Lombard effect, synchronously recorded with 3

The corpus includes 39 different, mainly male speakers with different accents of British English. Its main advantages are the realistic noise and channel characteristics, the availability of additional noise samples for noise simulation from the same domain and recorded by the same setup as well as the availability of synchronous recordings of the same utterances with both microphone technologies. As major disadvantages we can identify the small size of some of the evaluation sets, the lack of female speakers and the limited domain of command and control influenced by the British police forces.

The third corpus, the TETRA Corpus, is based on broadcast data and offers the possibility to test the approaches for a complex large vocabulary continuous speech recognition (LVCSR) system. It is one of the most complex tasks in ASR usually incorporating triphone acoustic models and complex language models learned from huge amounts of training data. The Fraunhofer IAIS LVCSR system detailed in [117] is based on a training set of currently almost 100 hours of speech (AM Corpus) with language models learned from texts of about 150M words. This corpus contains mainly planned clean speech from broadcast news. We further extended this data by various simulations and recordings introducing

certain TETRA radio channel characteristics to the clean speech. Simulations and recordings of TETRA radio transmitted signals enable a step by step estimation of the influences of several sources of degradation from high quality broadcast audio to highly compressed TETRA speech. The evaluation set only considers speech without background noise. The steps of simulation include low-pass filtering, AMR and TETRA coding of the signal and a realistic transmission via the TETRA channel using TETRA hardware. Its main advantages include the large amount of available training and test data, the complex and highly relevant task of LVCSR, the comparability of the influences due to the same baseline of speech data used for simulation and the actual transmission of speech via the TETRA channel providing realistic data for evaluations. The main disadvantage is the close to realistic but not realistic data as we use TV broadcast data transmitted via TETRA, which is atypical for real TETRA communications with, for example, a more spontaneous speaking style.

3.5 Summary

In this chapter we presented three corpora that we will use for various evaluations in the following two chapters. We summarised the main characteristics of each of the corpora and gave a more detailed description on the two newly created corpora, the TETRA Corpus and especially the purposely designed and recorded MoveOn Corpus. We further stressed strengths and weaknesses of the evaluation sets, the relevant types of distortion present in some or all of the data, and the ASR task for which the corpora are designed.

The TETRA Corpus and the MoveOn Corpus are particularly interesting in Chapter 4 about acoustic distortion analysing the effects of various sources of distortion introduced to the speech signal. All three corpora are finally used for evaluation of our approach of blind acoustic model selection and relative feature normalisation in Chapter 5.

Chapter 4

Acoustic Distortion

In Section 2.3 we presented major sources of acoustic distortion influencing the speech signal and the speech recognition accuracy in different ways. We analyse various of these sources from background noise, microphone channel characteristics towards hardware and transmission channel influences in view of automatic speech recognition (ASR) in this chapter. While evaluations and results on some aspects of acoustic mismatch and distortion as well as their influence on ASR are presented in various scientific papers, most publications do not offer extensive details or only focus on one specific aspect. We present an extensive evaluation on various sources of distortion in this chapter and also evaluate the differences between realistic and simulated noise. We analyse and understand the impact of realistic and simulated distortion on acoustic features and speech recognition performance and provide recommendations about promising ways to make an ASR system robust in various situations.

The evaluation corpora presented in the previous chapter — especially the MoveOn Corpus and the TETRA Corpus — enable us to evaluate several different sources of distortion and their effects on the speech signal and ASR separately. In an introductory section we present an exemplary integrated approach to build a robust speech recognition system for command and control on the motorcycle. We briefly present the different components relevant for the overall robustness of such a system including effects caused by hardware decisions, acoustic models, lexical knowledge and system restrictions.

As low-quality speech signals and mismatch between acoustic models and test utterances are found to be one of the most common problems, we go into detailed analyses of these aspects in the subsequent part of this chapter. In a first evaluation we will focus on the effects of background noise typically also referred to as additive noise. In addition to the distortion of the signal and speech features we address the problem of common noise simulation approaches usually assuming that speech and noise are uncorrelated. While often artificially noisy speech data is used for evaluating robust ASR approaches, we show that a realistic simulation of noisy speech data is hardly possible with common simulation approaches.

We further analyse channel effects causing variations and distortion in speech signals. First, two rather different microphone channels — throat microphone and close-talk microphone — are analysed and compared. Differences in spectral and cepstral representation of synchronously recorded signals from both channels indicate complex variations. In particular the differences in way and position of speech capturing influence the signal's quality as well as speech recognition accuracy in several ways. The influence of other channel effects from general hardware to speech transmission with coding/decoding effects is discussed in another section. A step by step evaluation of the different aspects of the TETRA radio channel separates various influences on the speech signal and their impact on the speech recognition performance. We show that especially harmonic distortion introduced by the hardware can severely influence the recognition accuracy.

The evaluations in this chapter provide an overview of the variety of distortion and their complex effects on speech features and ASR. Furthermore, we will show that even careful simulations of realistic distorted speech data usually provide an insufficient representation of realistic distortion. Complexity

and variability of acoustic distortion as discussed in detail in this chapter make the development of universal mismatch compensation and noise simulation methods very challenging and call for using realistic and non-mismatched data for training and evaluation whenever possible.

4.1 An Integrated ASR System

The performance of ASR depends on several factors, including the complexity of the task, signal variability and availability of representative training data. In an exemplary evaluation we will show by the design of a command and control ASR system based on the MoveOn requirements that a robust system with sufficiently reliable ASR performance is often possible, if the modules of an ASR system are well adapted to the purpose of the system. We will further evaluate the effect on the performance when replacing one or another of the modules with a less suited or alternative solution. The experiments will show the importance of a proper system layout and the strength of a properly adapted system with available adaptation data compared to a solution that must rely on mismatch compensation steps only.

The central aspects for achieving a reliable system, as discussed in the following section, comprise robust speech capturing, layout and adaptation of the system to the domain and purpose, training of well adapted acoustic models, and evaluation of additional mismatch compensation algorithms. The results show that a reliable command and control system for the task of command and control on a motorcycle is generally feasible. Our presented work was published in [3].

4.1.1 The Command and Control Task

The MoveOn multi-modal system is designed to provide communications and device control services to police motorcyclists on the move. The system provides several useful services to police officers which are cumbersome or impossible to use on the motorcycle without a well designed interface for human-computer interaction. The services include hands-free radio control, digital radio data transfer, camera control, a voice notepad, status management, and general system control. For example, the user is enabled to change radio channels by voice control, report his status and his position — acquired by a GPS device — to the police central, take pictures using a camera integrated in the helmet, store and retrieve information using a voice notepad, and change system parameters like the feedback volume. For our system design we focus on these tasks to derive command and control utterances to trigger and control such actions.

The main challenges for speech recognition for command and control in our work are the environmental conditions — mainly background noise and limited space on the motorcycle and in the helmet — as well as the requirements for a reliable system with minimal distraction to the user.

4.1.2 A General Integrated Approach

Today, an all-round solution for a robust speech recognition system that is working reliably even under all possible acoustic conditions with various sources of distortion as discussed in Section 2.3 does not exist. However, several aspects from speech capturing to speech processing can be taken into account to reduce the influence of some or all of the sources of distortion and to improve the overall robustness of the system in difficult environments. The implementation of signal processing algorithms to perform noise reduction on the noisy speech data is one popular approach. But as the influences of the various distortions are manifold, such an approach is usually not as successful as a full and integrated design of the entire speech recognition system. We will show and evaluate some of the central aspects to the system's robustness.

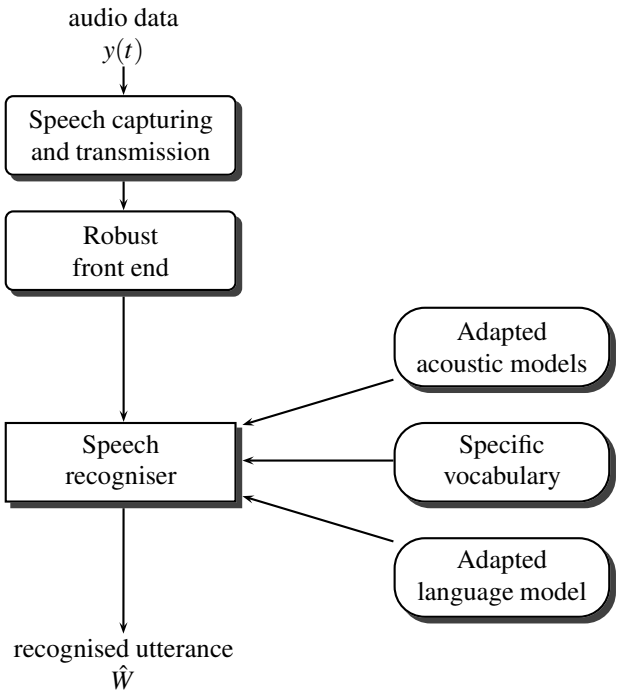


Figure 4.1: Components of an integrated robust ASR system. Several components of an ASR system influence the quality of the recorded speech signal and the speech recognition system. Designing all of these components from speech capturing and transmission, robust front end processing towards acoustic and linguistic knowledge sources appropriately improves the robustness of the overall system.

In Figure 4.1 the components of an integrated robust ASR system are pictured. As a first step towards robustness the recording hardware and transmission channel should be inspected to avoid unnecessary distortion due to unfavourable channel characteristics (compare 2.3.3). Here, a careful selection of appropriate hardware is essential to ensure a good quality of the speech data and to reduce environmental noise already during the process of capturing the audio signal. While we will show the general effects of using appropriate hardware here, we will further evaluate the influence of different microphone channels and low quality transmission channels in much more detail for two exemplary cases in Section 4.3 and 4.4. To further reduce noise and other signal distortions the signal quality can be improved by a preprocessing algorithm like a robust front end. Existing mismatch that could not be avoided in the first step should be reduced by equalisation methods (e.g. Section 2.4), while algorithmic distortion caused by these methods or the feature extraction should be avoided. Ideally, this block should output the exactly same features for the exactly same utterance no matter of speaker, environmental or any other acoustic mismatch. Practically, state-of-the-art feature extraction is still far away from this ideal output. The extracted features are finally processed by the speech decoder. The speech decoder itself also influences the performance of the recognition system dependent on the implemented algorithms. However, of much greater importance for the ASR results are the acoustic models, the language model (or grammar) and the definition of the vocabulary and pronunciation of the words provided to the decoder to avoid acoustic or domain mismatch in these knowledge bases compared to the input signal. In the following section, the components of the speech recognition process in Figure 4.1 are described in more detail, and the choices for our command and control system in our evaluation are provided.

Speech Capturing and Transmission

The quality of the recorded speech is a crucial factor for the performance of a speech recognition system. While the recording equipment for many speech applications cannot be influenced — e.g., for automatic transcriptions of broadcast data or for software consumer products for the mass market — the choice of an appropriate microphone and hardware setup can reduce channel distortion and improve the quality of the audio data significantly directly influencing the speech recognition performance. A good frequency response of the microphones in the relevant speech bandwidth is generally desirable, and a low distortion even for high acoustic pressure levels is important for loud and varying environments. Directional microphones capture speech coming from a certain direction and attenuate environmental noise and other speech from all other directions. Noise robust microphone technologies like active noise cancellation microphones or throat microphones are also interesting for noisy environments. Unfortunately, most of these noise robust mechanisms influence the microphone characteristics either in all cases or in case of inaccurate handling.

The transmission channel might also introduce significant distortion. A wireless connection, for example, often uses some speech or audio coding technologies that can affect the signal quality especially for low transmission bandwidth. Furthermore, certain disturbances and interferences might influence the electro-magnetic transmission introducing noise to the transmitted signal. Thus, wired connections are usually beneficial in terms of good channel characteristics with low distortion. The microphones used for capturing speech should provide a good frequency response and a sufficient dynamic range. Noise robust or noise cancellation technologies can be an option in noisy environments but can have negative effects on the frequency response. More details about the microphones used here are found in Section 3.3.3.

Robust Front End

In difficult signal-to-noise ratios, speech recognition performance usually falls off dramatically. Several approaches to signal preprocessing have been devised in order to counter this effect. While such approaches are not always beneficial and might introduce additional mismatch, improved recognition accuracies can be achieved in situations where unavoidable mismatch has to be considered. Some of the most common approaches were already introduced in Section 2.4.

In our experiments we will use an approach elaborated and standardised as part of the standard ETSI ES 202 050 [86].The robust front end of the standard implements several steps for noise reduction and blind equalisation and is explained in more detail in Section 2.4.5. Even though many scientific publications show certain improvements compared to the ETSI implementation, the ETSI robust front end opposed to most other approaches has proven to be successful to compensate several sources of acoustic mismatch in more than one specific setup. Thus, we include the ETSI robust front end in our experiments as an exemplary front end for mismatch compensation and robust feature extraction.

Speech Decoder

Available HMM-based speech decoders often use slightly different algorithms or approaches for calculating the best hypothesis of a spoken utterance. This mainly influences the decoding speed of a recogniser, which most developers try to optimise while the recognition accuracy is often considered as the parameter to keep constant close to the possible maximum. Thus, the difference in recognition accuracy between different popular state-of-the-art speech decoders is quite low and depends more on an expert settings of the parameters than the decoder itself. For the task of command and control a fast decoding is very important. As close to real time decoding for this rather simple ASR task is generally not a problem for most speech decoders, a fast decoding of speech provides more computing time for an effective robust front end for feature extraction. For the evaluation in this work, however, processing time is not evaluated. Thus, we do not consider the recognition speed of available recognisers and use the Hidden Markov Model Toolkit (HTK) ([37]), which provides all necessary tools for training, recognition and evaluation.

Vocabulary and Pronunciation

In small to medium vocabulary tasks like our task of command and control with a fixed set of commands, a careful selection of the commands and thus the vocabulary can increase the robustness of the ASR system. Mainly three characteristics of the vocabulary directly influence the performance of an ASR system: the size of the vocabulary, the phonetic distance between the words in the vocabulary, and a correct definition of the pronunciation of the words according to the targeted accent. Generally, a small and phonetically balanced vocabulary increases the robustness of the system, but limits the user experience and the field of application. In practice, a trade-off between functionality and reliability must be made.

Our speech recognition system for command and control is reduced to a rather small number of necessary commands with a focus set on reliability and less on versatile functionality. Whenever possible, we avoid very short words and words with short phonetic distances between each other. As the system is developed for British police forces, the pronunciation of the words is extracted from the BEEP dictionary for British English[1]. Our vocabulary contains 134 different words, including device names, command words, numbers and the international radio-telephony spelling alphabet.

[1] BEEP Dictionary - http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

Device	Command	Parameter
Status	set	patrol | pursuit | emergency
Radio	change channel	[call sign]*
	change folder	[call sign]*
Camera	open	—
	close	—
	take	picture
Volume	up	—
	down	—
	mute	—
—	confirm	—
	deny	—
[call sign] = [Alpha-Zulu] [Alpha-Zulu] [0 - 99]		

Table 4.1: Examples of MoveOn commands. Most commands defined and recorded for the MoveOn Corpus are defined by specifying a device to control, a command to trigger a certain action, and a parameter specifying a new value or state. Some typical commands are presented in this table.

Language Model vs. Grammar

The language model (or grammar) of the system makes use of higher level syntactic relations in utterances. The adaptation of the language model to a certain domain usually improves the recognition results [118, 119] as probabilities of specific syntactic relations are also dependent on the domain. A statistical language model is more adaptive than a grammar (compare Section 2.2.3), but for a robust ASR system for command and control, a fixed definition of word orders in such a grammar can be of advantage to increase the robustness and reliability of a system, as long as the users formulate their utterances accordant to that grammar.

Thus, we define a finite grammar for the the task of command and control. The grammar includes relevant commands for human-computer interaction systems for police motorcyclists and is based on the AW and CP items of the MoveOn Corpus (Sections 3.3.3 and 3.3.5). Most phrases defined by the grammar can be described by the following structure:

[device] <command> [parameter]

Essential commands like "confirm" or "cancel" contain one word only. Usually, commands specify the device to be controlled and the command to trigger a desired action for this device. Furthermore, some commands require a parameter to define a new value, e.g. "radio + change channel to + two" (<device> + <command> + <parameter>). Thus, every item in the structure above can contain more than one word. We chose this grammar to enable an intuitive system, which is still restricted enough to enable a reliable recognition performance in the difficult environment on the motorcycle. The grammar incorporates 134 words and has 1016 nodes in the graph representation. Its perplexity is about 7.3 (compare Section 2.2.3). The command and control phrases have different lengths between 1 to 11 words with an average of 3 words.

In Table 4.1 some examples for MoveOn commands according to this grammar are listed. Parameters are only necessary for specific commands. Some fundamental commands — especially for confirmation or rejection — are used without any device specifier.

	Clean	Noisy
Close-talk microphone	1700	5500
Throat microphone	1500	4100
VoxForge	*n/a*	*n/a*

Table 4.2: Amount of available training data in number of utterances. For the evaluation on the integrated robust ASR system the following number of utterances were available for training acoustic models for each acoustic domain. The VoxForge acoustic models were used as provided for comparison. The amount of training data available to train these acoustic models was not specified in detail.

Acoustic Modelling

Acoustic models are a central part of a statistical approach for speech recognition. The smaller the mismatch between the data used for training the acoustic models and the data seen during recognition, the better the recognition results.Training or adaptation of specialised acoustic models is probably the most important aspect of creating a robust speech recognition system as also shown in, e.g., [120]. While training of completely new acoustic models from scratch is usually the best approach to yield good ASR results, a rather large amount of transcribed speech data from the target acoustic domain is needed. Thus, adaptation techniques are often applied to adapt existing acoustic models trained on large amounts of data from a similar acoustic domain or speaker (compare Section 2.4.7).

In our experiments we train new sets of acoustic models from the data of the MoveOn corpus, each specialised for one target acoustic domain. Considering two different sources of possible mismatch, we train four sets of acoustic models for all combinations of the two available microphone channels (throat microphone and close-talk microphone) and the two environmental conditions (clean office speech and noisy motorcycle speech). Thus, we can show the effect of the different microphone technologies in case of matched and mismatched conditions for a realistic scenario of command and control on the motorcycle.

4.1.3 System Evaluation

In these preliminary evaluations we test the impact of the various components of a robust ASR system on the recognition accuracy. Four major aspects are considered in detail: the differences and effects of using two rather different microphone technologies, the significance of well-trained or adapted acoustic models, the effect of the finite grammar and small vocabulary, and the influence of a robust speech recognition front-end in case of matched and mismatched conditions.

Microphones and Acoustic Models

As acoustic features inevitably contain information about the acoustic conditions from speaker and speech variability to channel characteristics, acoustic models trained on such features will also contain this information causing mismatch and dropping recognition accuracies for data with other acoustic conditions (compare Section 2.3). To evaluate the difference between matched and mismatched situations and the influence of the microphone channel on the ASR performance, we train four acoustic models for all combinations of acoustic environment and microphone channel. The impact of the environmental mismatch is investigated for each microphone channel by comparing the performance of the system using acoustic models based on data of the realistic (noisy) environment, on the one hand, and acoustic

Speaker	Gender	Environment	Commands
s1	female	office (clean)	93
s2	male	office (clean)	73
s3	male	motorcycle (noisy)	60
s4	male	motorcycle (noisy)	81
s5	male	motorcycle (noisy)	86
all			**383**

Table 4.3: Amount of available test data per speaker in number of commands. The number of utterances (commands) available for each of the test speakers $s1$ to $s5$ for evaluation of the integrated robust ASR system are listed. Additional information about gender of each speaker and recording environment is provided. The test utterances are available for both throat and close-talk microphones.

models based on data recorded in a silent office environment, on the other hand. Furthermore, the results are compared to results based on VoxForge[2] open source models, providing acoustic models from a completely different domain, with presumably significant hardware and environmental mismatch.

The MoveOn acoustic models are trained from the data of the MoveOn Corpus (Section 3.3) downsampled to 16 kHz at 16 bit. The experiments in this section were done before the MoveOn Corpus was finalised, so a different training and test set splitting compared to Section 3.3.5 was used here.We train two acoustic models for each microphone channel, a set of clean speech models, which are based on the silent office data of the MoveOn Corpus, and a set of noisy speech models, which are based on the MoveOn Corpus data recorded in the realistic environment on the motorcycle. Our test set contains two speakers (one female $s1$, one male $s2$) from the office subset and three speakers (all male, $s3$-$s5$) from the motorcycle subset. The exactly same speakers are used in the evaluation of both microphone channels. Only utterances in compliance with the command and control grammar definition are included. The number of test utterances per speaker is listed in Table 4.3. The data of all other speakers is used for training. The amount of available training data for each set of acoustic models is listed in Table 4.2. As we use all available data of each microphone channel the overall amount of data for the throat microphone is generally smaller than the available amount of data for the close-talk microphone.

We use our standard setup of MFCCs as speech features. Twelve coefficients with additional short time energy and first and second derivatives are calculated and cepstral mean normalisation is performed. Due to the small to medium amount of training data we do not model the acoustic context and use simple monophone models instead of triphones with 16 Gaussian mixtures per state. We use HTK to extract the speech features, except for the ETSI robust front end, where we use the front end to extract 13 static ETSI features together with their first and second order derivatives. Comparable acoustic models are trained from each training set using HTK.

For the evaluation of the acoustic models of the throat microphone the throat microphone test data is used. All other models are tested with the close-talk microphone data.

We determine the word accuracy rate to evaluate the overall quality of the command and control system. In Table 4.4 we can see the word accuracy rate for each channel and test speaker based on our evaluation set for command and control. The test sessions recorded in the realistic environment — including utterances with low SNR — show a very good performance with an accuracy rate in the range of 95% to 100% for matched conditions. The recognition performance for the throat microphone channel on matched conditions is slightly lower. For mismatched conditions due to the environment the

[2] VoxForge — OpenSource data for speech recognition: http://www.voxforge.org

	Clean		Noisy		
	s1	**s2**	**s3**	**s4**	**s5**
Close-talk models, clean	98.15	**100.00**	66.13	93.80	68.24
Close-talk models, noisy	81.18	92.27	**100.00**	**100.00**	**98.82**
Throat mic. models, clean	95.94	99.55	96.77	95.04	78.43
Throat mic. models, noisy	20.30	97.73	99.46	100.00	96.47
VoxForge models [121]	79,34	83.64	46.77	80.17	52.55
ETSI close-talk models, clean	**98.52**	99.09	88.71	**100.00**	86.67

Table 4.4: Word accuracy rate for different acoustic models and test speakers. Acoustic models with matched and mismatched conditions are evaluated with the test data of the five speakers. Throat microphone acoustic models are evaluated on the throat microphone test sets, all other acoustic models are evaluated on the close-talk microphone test sets.

throat microphone channel is less affected and performs almost equally well on clean and noisy speech utterances for both sets of acoustic models (except Speaker $s1$ and $s5$ with more considerable mismatch), whereas the close-talk microphone shows significant differences for all speakers with environmental mismatch. For female speaker $s1$ a very low word accuracy of only 20.3% in case of noisy throat microphone models can be seen. This is due to the small neck of the female speaker causing a lack in pressure of the throat microphone transducer on the larynx. Including another female speaker in the training set of the clean throat microphone models compensate for some of the mismatch in the features introduced by the lack of pressure improving the results to almost 96% word accuracy. As we will see later, this difference in word accuracy of over 75% absolute between both models is smaller (about 40% absolute difference) for the acoustic performance without grammar influence (compare phoneme and word accuracy rates in Figure 4.2).

In mismatched conditions, when acoustic models from a different environment are used, the results drop significantly. In particular when using the open source models from VoxForge, which mainly contain speech recorded with random desktop microphones in an office environment, providing significantly lower results for the clean speech test data. This is presumably caused by a mismatch of the recording equipment and room characteristics (compare Figure 2.8). The performance of the VoxForge acoustic models is even lower for the data recorded in the noisy environment as the background noise causes additional mismatch between acoustic models and test data.

Comparing the performance of the acoustic models trained on the standard MFCCs of the clean speech data of the close-talk microphone with the acoustic models trained on the features extracted by the ETSI robust front end show the capabilities and risks of applying such a robust front end. While the word accuracies for the noisy speech test data significantly improve for all speakers, the word accuracies for the clean speech test data only slightly improve by 0.5% absolute for the female speaker and decreases by about 1% for the male speaker. We see that the front end is able to reduce existing mismatch but can cause algorithmic distortions and a slightly decreasing performance if hardly any mismatch is present. In spite of the improvements in mismatched situations, the performance of the ETSI robust front end is still below the performance of the acoustic models from the same acoustic domain as the test data avoiding any significant mismatch.

Evaluation of the Finite Grammar

The effect of a small vocabulary and a finite grammar is evaluated by comparing phoneme recognition results with word recognition results. For an estimation of the influences of defined syntactic relations

	Clean		Noisy	
Throat mic.	**Phoneme**	**Word**	**Phoneme**	**Word**
s1	47.46	95.94	12.12	20.30
s2	72.81	99.55	50.36	97.73
s3	43.42	96.77	75.00	99.46
s4	46.65	95.04	62.64	100.00
s5	29.68	78.43	50.32	96.47

Table 4.5: Comparison of phoneme and word accuracy rates. For the throat microphone evaluations from Table 4.4 the average phoneme and word accuracy rates for each speaker are listed. The finite grammar used during sentence recognition is capable of compensating for some phoneme recognition errors until the word accuracy also drops significantly for very low phoneme accuracy rates.

two recognition processes are performed. First, we use an ASR system for recognition on a word level including the task specific syntactic relations defined in the vocabulary (phoneme relations) and finite grammar (word relations) described in Section 4.1.2. Second, we set up an ASR system that only performs a phoneme based recognition without high-level semantic relations, i.e., the vocabulary for the recognition only contains the phonemes with the assumption that all probabilities and conditional probabilities of the phonemes are equal. So just the acoustic information is taken into account for this recognition process.

In Table 4.5 the effect of the limited 134 words vocabulary and our finite command and control grammar is illustrated by comparing phoneme and word accuracy rates for two examples showing typical value pairs for phoneme and word accuracy of an utterance. While, for example, the phoneme recognition rates for speaker *s2* and *s5* for the acoustic models of noisy throat microphone speech are much lower than the rates for *s3* and *s4*, the word recognition rates are almost equal and close to the maximum. For very low phoneme recognition rates (e.g. *Noisy*, *s1*) the word recognition rate drops significantly. In Figure 4.2 this effect is plotted for a larger number of values: the finite grammar is capable of improving phoneme accuracy rates down to about 40% still achieving satisfying word accuracy rates of 90% and more. The additional lexical knowledge provided by the command and control grammar was able to correct several of the phoneme errors of the acoustic recognition process as long as a sufficient number of correctly recognised phonemes is available.

Microphones and Robust Front-End

We already briefly discussed the influence of the ETSI robust front end and the microphone channels on the word accuracy. Now we want to have a closer look on the effects of the noise robust throat microphone channel as well as the ETSI robust front end.We now compare the performance of 6 different acoustic models — a clean speech and a noisy speech set of acoustic models for standard MFCCs of the throat microphone and the close-talk microphone channel as well as for ETSI features of the close-talk microphone channel. All test data is based on the realistic noisy recordings from the motorcycle environment. As we do not consider any grammar here, we also include all other noisy test utterances that do not follow the command and control grammar. This results in a test set of 588 utterances. The type of features and the microphone channel is always matching in this evaluation, only the environmental conditions might be non-matching. We evaluate the phoneme accuracy to compare the performance of the acoustic models and features without incorporating any lexical knowledge or grammar. The test data is grouped by different levels of signal-to-noise ratio (SNR) estimated by the NIST STNR tool of

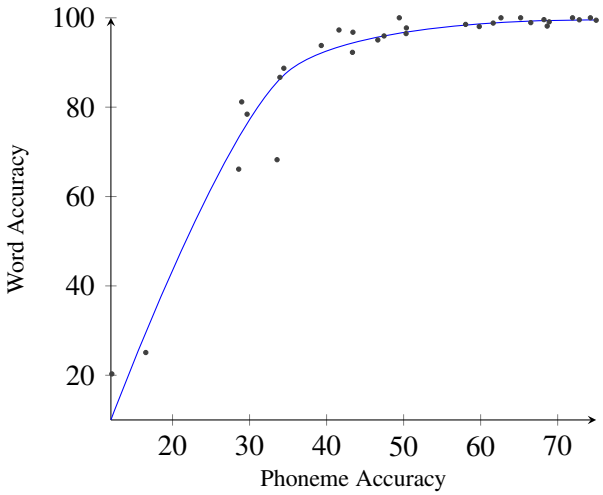


Figure 4.2: Dependency of word accuracy rates from phoneme accuracy rates. The graph shows the dependency of the word accuracy rate from the phoneme accuracy rate when using the finite grammar. The results from both microphone channels are combined. While some phoneme recognition errors can be compensated due to the restrictions of the finite grammar, word accuracy drops significantly if the phoneme accuracy rate gets too low.

SNRs:	**all**	**>20dB**	**15-20dB**	**10-15dB**	**5-10dB**	**<5dB**
Throat mic. models, clean	35.53	37.43	38.05	34.05	34.26	29.73
Throat mic. models, noisy	53.89	54.60	55.96	53.84	54.06	44.46
Close-talk models, clean	30.72	37.24	33.91	27.12	26.78	23.70
Close-talk models, noisy	**58.45**	**61.44**	**61.95**	**59.03**	**54.81**	**46.42**
ETSI models, clean	36.11	39.25	38.42	34.99	33.82	29.31
ETSI models, noisy	56.23	58.29	58.44	57.02	54.61	45.02

Table 4.6: Phoneme accuracy rates for different SNRs. The evaluated acoustic models are trained on the clean speech or noisy speech training data of each microphone channel and for the close-talk channel with ETSI robust feature extraction. The test data is noisy speech from the same microphone channel and front end as the acoustic models.

the NIST Quality Assurance Package[3]. The test data of the throat microphone channel is grouped by the estimated SNR of the corresponding close-talk microphone utterance as we want to compare the performance of both microphones under different levels of environmental noise. These conditions are much better represented in the close-talk microphone signal.

Table 4.6 shows the phoneme accuracies for the different SNRs and acoustic models. The best results are achieved by the acoustic models trained from data of the same environment, i.e., also trained on noisy speech data. Without any mismatch, the ETSI front-end cannot improve the recognition results — the phoneme accuracy is even slightly worse. But for mismatch between training and test data, the ETSI front-end is able to significantly improve the recognition results. It improves the phoneme accuracy for mismatched clean speech acoustic models by 2 to 7% absolute (5 to 30% relative improvement).

Figure 4.3 visualises the results from Table 4.6. The mismatch of the acoustic models trained on clean speech data to the realistic data in a noisy environment leads to poor recognition rates for the close-talk acoustic models trained on clean speech data — especially for low SNRs. The throat microphone is less

[3] NIST SPQA 2.3, `http://www.itl.nist.gov/iad/mig/tools/`

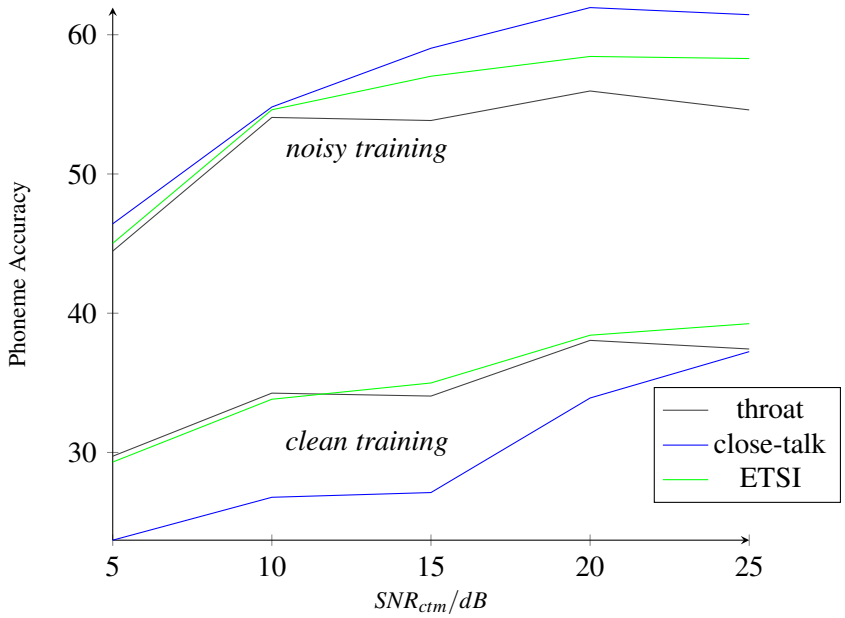


Figure 4.3: Phoneme accuracy rates for noisy speech on different evaluation setups. The results of Table 4.6 for noisy test data evaluated on different sets of acoustic models and front ends are pictured. Throat microphone and ETSI front end trained on mismatched clean conditions improve the results compared to mismatched close-talk acoustic models. Non-mismatched conditions providing noisy close-talk acoustic models perform best.

affected by the environment due to its particular speech capturing mechanism, and therefore, achieves much better results when trained on clean speech data but tested in noisy environments. If a mismatch between training data and recognition data cannot be avoided, the ETSI front-end promises to improve the results. Still for acoustic environments, where the noise conditions are not too manifold and varying, the training of acoustic models from data recorded in exactly the same environment is usually a much better solution.

4.1.4 Conclusion

In this evaluation the different aspects of developing a noise robust speech recognition system for a command and control system on the motorcycle were discussed and major components and concepts relevant for the robustness were evaluated. It is recommended to tune every component of the system to achieve a system as robust as possible. In particular the acoustic models should be well-trained on or adapted to the target acoustic conditions. A robust front end is particularly interesting, whenever mismatch between training and recognition data cannot be avoided. Finally, we showed the impact of a finite grammar to make our system more robust providing good word error rates even for lower phoneme accuracies. While this can be an interesting option for certain tasks, this aspect is not applicable in all cases. In case of a LVCSR system we cannot avoid to use a non-restricted language model and a large vocabulary, so that much attention must be paid to properly adapt this language model to the statistics of the domain. In [122], for example, the authors show that such an adaptation also significantly improves the overall ASR performance.

4.2 Background Noise

In the previous section we could see the influences of the different components on the speech recognition performance for command and control. One major problem was mismatch in the acoustic features between acoustic models and recognition data caused by distortions of a speech signal as we already discussed in Section 2.3. Background noise is one typical type of distortion also present in the MoveOn Corpus with severe influences on the recognition accuracy as could be seen in the previous section. Thus, we want to have a closer look on the effect of background noise and in particular on the assumption that it can be described by an additive term in the time or frequency domain. Parts of the following work were already published in [5].

Much work on robust speech recognition focuses on background noise in particular. Often the assumption of independence is used to estimate and remove the noise from the speech signal (see Section 2.4.2). On the other hand, several evaluation corpora for evaluating such noise reduction approaches exist and are often created artificially (including Aurora 2, Section 3.1), i.e., noise is mixed with clean speech utterances in different SNRs assuming an additive mathematical model. Thus, such corpora provide artificially noisy speech which actually fulfils the assumption of independence of noise and speech used by the noise reduction approaches. But for real-life data this would not necessarily be true.Thus, we decide to investigate, how the performance of speech recognition and robustness approaches differ when comparing real noisy speech with very similar artificial noisy speech. During this evaluation we will further have a look on the general effect of additive noise on speech, speech features and recognition accuracy.

4.2.1 Related Work

Hardly any work directly focusses on differences between simulated and realistic background noise. Still, simulated noisy speech is widely used to evaluate algorithms for robust ASR. Most of the algorithms for noise reduction like the large group of spectral subtraction approaches based on [52] assume that background noise is additive. On the other hand, most evaluation databases with noisy speech are artificially created by adding some background noise recorded in a certain noisy environment to a clean speech signal in various SNRs. Examples for such databases are Aurora 2 (Section 3.1) and the HIWIRE database [123]. In Aurora 2 various samples from noisy environments like car or exhibition hall are added to the clean speech TIDIGITS database [102]. In the case of HIWIRE a noisy cockpit is simulated by adding noise samples recorded in an actual cockpit to non-native English speech from a clean environment. The application of ASR focused on in the HIWIRE database is command and control.

So while noise simulation has been used in a lot of work related to ASR, to our knowledge no detailed evaluation has been done so far to show how realistic such simulated noise actually is. Some work focussing on the Lombard effect in speech uttered in noisy environments (e.g. [64–66]) indicate that certain effects present in realistic environments might not be considered when just adding noise to clean speech. To this end we decided to investigate the differences between simulated and realistic additive noise and to evaluate the influence of background noise on speech and speech features at the same time in the following section.

4.2.2 Additive Noise Theory

Usually, background noise $n(t)$ is considered to be independent from a speech signal $s(t)$ and, thus, is modelled as additive noise as follows (with convolution $*$ and transfer function $h(t)$ — also compare to

Equation 2.30):

$$x(t) = [s(t) + n(t)] * h(t) \tag{4.1}$$

This simplified model as also shown in Figure 2.7 already neglects certain aspects compared to the complex model in Figure 2.8, for example the room effects influencing noise and speech before captured by the microphone. Thus, considering the complex model in the second figure, $s(t)$ in Equation 4.1 must be replaced by $\tilde{s}_m(t) = s(t) * h_{r,sm}(t)$ and $n(t)$ by $\tilde{n}_m(t) = n(t) * h_{r,nm}(t)$. In case of different rooms or microphone locations when recording speech and noise separately, this difference already introduces some error in the artificial noise signal compared to realistic conditions. For simplicity we use $s(t)$ and $n(t)$ here for the modified speech and noise signal on the location of the microphone, where both signals are finally mixed. Then, in case of a captured noisy speech signal, the same microphone and hardware channel characteristics can be assumed as both sources are captured simultaneously. Thus, we can describe the recorded noisy speech signal $x(t)$ by an additive combination of the recorded speech signal $\hat{s}(t)$ and the recorded noise signal $\hat{n}(t)$:

$$x(t) = \hat{s}(t) + \hat{n}(t) \tag{4.2}$$

with

$$\hat{s}(t) = s(t) * h(t), \tag{4.3}$$
$$\hat{n}(t) = n(t) * h(t). \tag{4.4}$$

In recordings of realistic noisy speech this assumption of independence is not generally correct, as a speaker in a loud environment with background noise is speaking in a different way than usual (compare Section 2.3.2). The effect is introduced to make speech in noise more intelligible to the listener. But it also means that speech is indeed dependent on environmental noise and the assumption that background noise is just additive is wrong as indicated by the connection of $n(t)$ and Speech Production in Figure 2.8. In case of noisy speech simulations, where noise and speech is even recorded in different setups, also the room characteristics as dicussed above might be rather different and not realistic. Furthermore, the assumption of an identical transfer function $h(t)$ in Equation 4.2 is usually wrong as soon as a different microphone, recording hardware etc. is used. Still, most simulation approaches just add any recorded noise to any recorded speech basically applying this equation.

4.2.3 Simulation of Additive Noise

We now want to compare realistic noisy speech with simulated noisy speech which should be as similar to the realistic noisy speech as possible. As we just discussed before, Equation 4.2 assumes that $h(t)$ is identical for recorded noise and speech. This is basically true for realistic noisy speech (ignoring smaller acoustic effects like near field acoustics etc.), as both signals are captured by the same equipment at the same time. The propagation path of speech practically does not change and the one of noise outside the helmet can be considered to be very similar for most recordings due to similar free field conditions and the same helmet setup. But when mixing speech recorded in one setup in a certain environment and noise separately in another setup in probably another environment, this assumption is already wrong and the equation becomes only an approximation. In the case of the MoveOn Corpus, speech and noise are recorded with exactly the same hardware inside a motorcycle helmet, so that both — recording channel and room characteristics — can be assumed to be practically identical for noise in pure noise and noisy speech recordings and for speech in clean and noisy recordings. Further assuming time invariance, the

same $h(t)$ can be assumed and the right-hand side can be used to just add the recorded noise $\hat{n}(t)$ to the recorded speech $\hat{s}(t)$ as shown in Equation 4.2. Hence, we have rather ideal conditions for a realistic simulation as opposed to most cases of simulated noisy speech data

Data Splitting

We divide the speech data of the right microphone channel of the MoveOn database into several subsets. In a first step all noisy speech data, i.e., all utterances that have any tag in the noise annotations (compare Section 3.3.4), is put in one cluster and the clean speech data (all other data from both office and motorcycle recordings) is put in another cluster. Please note that this is different to the data splitting of the MoveOn Corpus baseline experiments in Section 3.3.6. Additionally, we use all available noise segments from the motorcycle recordings without speech. Both — noisy speech and noise samples — are tagged with information about the types of background noise, which we will use to create simulated data close to the realistic reference data. As we have more than twice as many noisy speech utterances than clean speech utterances, we further split the noisy speech data into two reference noisy speech sets with different speakers. Due to the rather limited amount of data we use cross-validation iterating the six major speakers per set using one of the speakers for testing and all other speakers of the set for training in each iteration.

Simulation Process

Based on the two sets of reference noisy speech data we create a similar artificially noisy set based on the clean speech utterances and the noise samples. Therefore, we consider the following aspects to match reference and simulation as well as possible:

1. The phoneme sequence of the clean speech should be as similar as possible to the reference noisy utterance.

2. The noise sample should have exactly the same annotation (same noise in the background) as the noisy speech reference.

3. The resulting estimated SNR of the simulated noisy speech should be as close as possible to the estimated SNR of the reference noisy utterance.

Based on these three aspects we choose for each reference noisy speech file a clean speech file with a phonetic transcription as close as possible in terms of phoneme accuracy (Section 2.5.1, Equation 2.52) compared to the reference data transcription. This value is close to 100% in most cases, as the number of different utterances and the overall size of vocabulary of the corpus is rather small. Each clean speech file is only used once for simulating an utterance of the reference set. Next, we choose a noise sample exactly matching the noise annotation of the noisy speech file. Furthermore, the length of the noise file must be equal or larger to the length of the clean speech file to make sure that the resulting simulated noisy speech is distorted in full length.

SNR Adaptation

After clean speech and noise sample for each reference utterance are selected, the SNR of the simulated noisy speech is adapted to match the estimated SNR of the reference speech.

We can exactly calculate the SNR for the simulated noisy speech but not for the realistic reference speech as we do not have speech and noise separately. Thus, we aim at an identical estimated SNR for reference speech and simulated noisy speech.

For SNR estimation we use the *NIST segsnr* tool from the *NIST Quality Assurance Package*[4]. *NIST segsnr* requires a voice activity file specifying segments of speech in an utterance. As we have fully orthographic transcriptions of all utterances, we use acoustic models trained on all available training data (full set, right channel) defined in Section 3.3.4, Table 3.11, and forced alignment provided by HVite of HTK to cluster each utterance into speech and non-speech segments decided on speech monophones versus non-speech monophones (silence and short pause). As we use additional knowledge compared to a blind estimation in combination with well adapted acoustic models we expect improved estimations of the voice activity even for noisy speech.

To achieve a high degree of similarity between the estimated SNRs of the reference set and the simulated set, we first mix clean and noisy speech with the estimated SNR from the reference noisy speech utterance. As we exactly calculate the SNR during the mixing process, we assume at this step that estimation and calculation show the same results. Now we estimate the SNR of the resulting simulated noisy speech the same way we did for the reference noisy speech. Figure 4.4 shows the results. As we can see estimated and calculated SNR do not exactly match. Towards $5dB$ the estimation of the SNR also begins to level out due to imprecise estimations. Thus, we estimate a function for deriving the SNR to use in the simulation (SNR_{corr}) to achieve a desired estimated SNR of the resulting simulated signal (SNR_{est}). This dependency is approximated by a stepwise polynomial function that we can use to correct the values:

$$SNR_{corr} = a_3 SNR_{est}^3 + a_2 SNR_{est}^2 + a_1 SNR_{est} + a_0 \tag{4.5}$$

with

$$\begin{array}{lllll}
a_3 = 0, & a_2 = 0, & a_1 = 0, & a_0 = 0, & \text{for} \quad SNR_{est} < 6dB \\
a_3 = 0.0221, & a_2 = -0.8716, & a_1 = 12.269, & a_0 = -44.742, & \text{for} \quad 6dB \leq SNR_{est} \leq 10dB \\
a_3 = 0, & a_2 = 0, & a_1 = 1.1733, & a_0 = 1.1071, & \text{for} \quad SNR_{est} > 10dB
\end{array}$$

With Equation 4.5 we can calculate a corrected SNR to use (SNR_{corr}) from the estimated SNR (SNR_{est}) of the reference data. SNR_{corr} is then used during simulation to weight clean speech and noise sample in a second mixing step. In Figure 4.4 we can see that the correction improves the matching of estimated SNR of the reference data and estimated SNR of the simulated data. The resulting simulated data is used for the experiments to compare simulated noisy speech with very similar characteristics to the reference noisy speech.

Even though the SNR estimation and the selection of noise is not optimal, we can assume that we are much closer to an optimal simulation than most other simulated databases, which suffer from the same problems and additionally from different transfer functions caused by different environments and rooms.

4.2.4 Influences on Speech Characteristics

Figure 4.5 shows the spectrograms of the same utterance “Helmet Cam” (in SAMPA phonemes “h e l m I t k { m”) spoken by two speakers in clean (*a* and *c*) and noisy (*b*) or simulated noisy environment (*d*). The simulation of the noisy speech in (*d*) is based on the reference in (*b*).

[4] NIST SPQA 2.3, http://www.itl.nist.gov/iad/mig/tools/

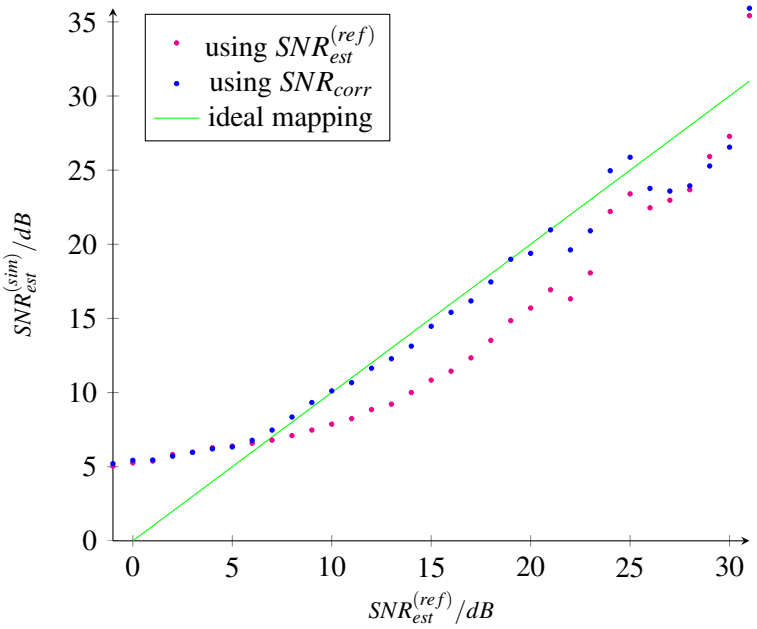


Figure 4.4: Correction of estimated SNR for noise simulation. A corrected value SNR_{corr} is determined for simulation to achieve that the resulting estimated $SNR_{est}^{(sim)}$ matches the estimated $SNR_{est}^{(ref)}$ of the reference. The graph shows $SNR_{est}^{(sim)}$ in case of a simulation with $SNR_{est}^{(ref)}$ directly and with an adapted value SNR_{corr}.

In the clean speech spectrograms we see similarities on the shape and development of the spectral peaks with time. But we also see many differences in speaking style between these two speakers in clean conditions. Speaker 1, for example, is speaking slower than speaker 2 and the energy is more concentrated on lower frequencies. The fundamental frequency of speaker 1 seems to be slightly higher than for speaker 2. In the noisy speech examples, strong engine and air wind noise were added in (*d*) at the same estimated SNR as for reference speech in (*b*). The effect of the noise can easily be seen in the spectral noise. Except for the added noise, the speech spectrogram in (*d*) is not modified as we can see when comparing it to (*c*). In (*b*) on the other hand we have realistic noisy speech from speaker 2 in strong engine and air wind noise conditions. Compared to the clean speech of the same speaker in (*a*) we can see various differences also in the speech characteristics. While the tempo seems to be similar, a shift of the energy from low to medium frequencies for the vowels /*e*/ and /{/ and from low to higher frequencies for the change from /*I*/ to unvoiced stop /*t*/ can be seen. This is in line with the possible changes caused by the Lombard effect as described in Section 2.3.2, while some other possible effects cannot clearly be recognised. Still, this is just an example showing that there seem to be some influences on the speech caused by environmental noise, which are not present in the simulated speech data.

In Figure 4.6, left-hand side, we compare the spectrum over the whole utterance and exemplarily the third cepstral coefficient for the utterance for all 4 examples. For the spectrum of the first speaker we can see that a simulation of noisy speech fills the valleys in the spectrum but hardly changes the peaks. Particular patterns in the simulated noisy file are comparatively hard to discover. The clean speech sample of the second speaker also shows particular patterns but the peaks in the spectrum are at different frequencies. The main reason are the speaker's differences in particular caused by the different characteristic frequencies of each speaker. The clean speech utterance and the noisy speech utterance of the second speaker are still rather similar. While the valleys are also filled up here by the noise patterns, certain peaks are still rather distinct and the speech energy seems to be distinctively higher than for

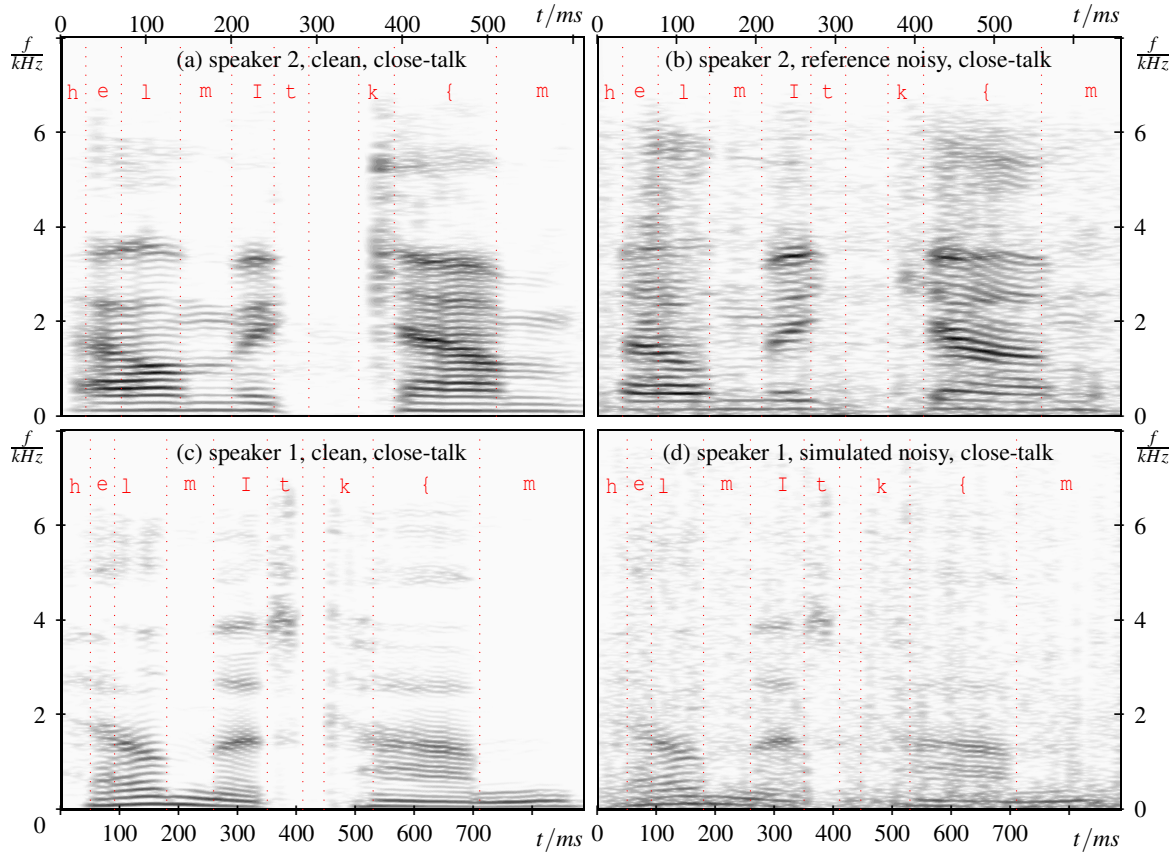


Figure 4.5: Spectrograms for clean, simulated noisy and realistic noisy speech. The spectrograms show the utterance "Helmet Cam". (b) is realistic noisy speech used as reference for simulation. (d) is the simulated noisy speech based on clean speech (c) and noise characteristics of (b). (a) gives an example for clean speech from the same speaker as the reference noisy speech (b).

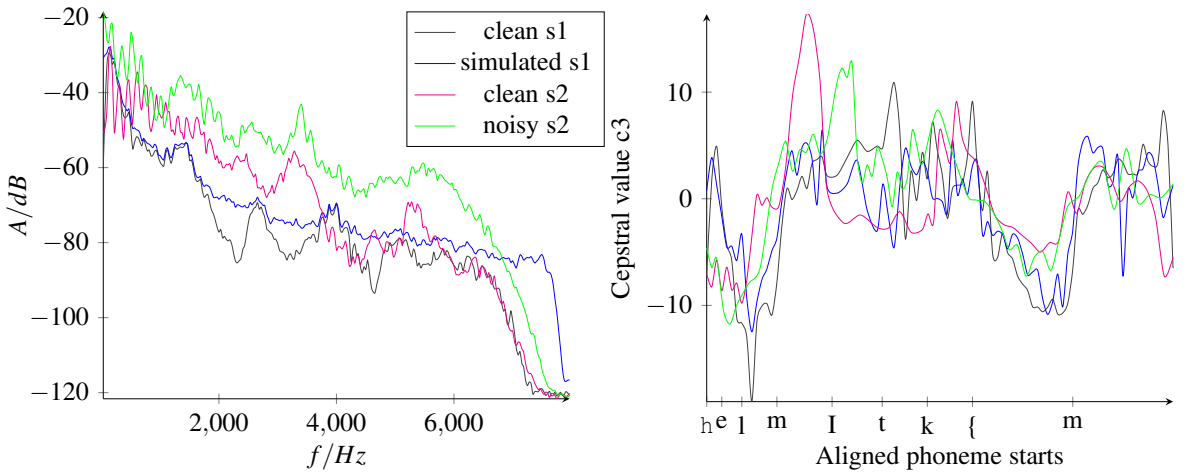


Figure 4.6: Example for influences on spectrum and cepstral values caused by speaker and noise. The figure on the left-hand side shows the spectrum for the utterance "Helmet Cam" for the examples from Figure 4.5. The figure on the right-hand side compares the aligned third cepstral coefficients for the same examples.

clean speech.

When we compare the spectrum of clean speech, simulated noisy speech and reference noisy speech in Figure 4.7 for the phonemes /{/ and /m/, we can see the same effect of filled valleys for simulated speech. Even thought the signal becomes more similar to the reference noisy speech, we can see several differences for both phonemes, in particular a lack of clear peaks in the spectrum (which are probably emphasised by the Lombard effect in the realistic case). Thus, simulated noisy speech does not seem to be too similar to reference noisy speech even in the case of our rather controlled conditions. We further can expect that simulated noisy speech will perform even worse in ASR than noisy speech, as clear peaks are missing, which help to identify the phonemes.

Finally, we compare the cepstral coefficients in Figure 4.6, right-hand side. We aligned the third cepstral coefficient[5] based on the phoneme alignments of HTK. Thus, we have the time warped positions of comparable cepstral values for all four example recordings. In general, all four examples show roughly the same progression of this cepstral coefficient, but the variations around the common trend are rather high. For speaker 1 we can see that for clean speech the third cepstral coefficient is showing distinct peaks, while the simulated version shows different characteristics for many phonemes of the utterance. The difference of the cepstral coefficient can be even more than 10 units, which is about one third of the observed range of the third cepstral coefficient in the whole utterance. Comparing both examples for speaker 2 reveals even slightly higher differences in the cepstral values of the third coefficient especially for the phonemes /m/ and /I/. In general, the variety of this cepstral value is very high even for clean speech but two different speakers. Additive noise and other effects seem to create even more variations in the cepstral values. Other cepstral coefficients usually show similar trends for different speakers or acoustic conditions.

4.2.5 Evaluation of ASR Performance

We could see in the previous section that several effects including additive noise seem to affect the cepstral values used for ASR. If we assume that we are able to simulate the noisy speech data very

[5] The selection of the third coefficients is arbitrary. The general effects discussed here are comparable for the other cepstral coefficients.

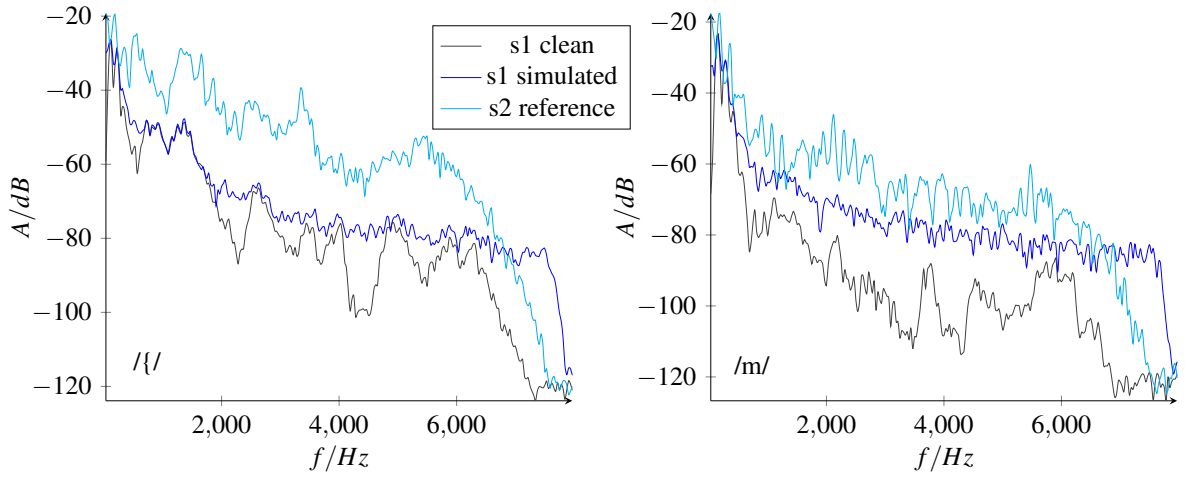


Figure 4.7: Influence of the simulation on the spectrum of the phonemes /{/ and /m/. The spectra for both phonemes show various differences in the three signals. In general background noise fills "valleys" in the spectrum hiding peaks of the speech signal. Speaker differences between speaker $s1$ and $s2$ are also visible by different positions of the peaks.

well using an additive noise model, simulated and realistic noisy speech should be rather similar. If we further consider the results from Section 4.1, acoustic models trained on one of the sets (either reference noisy speech or simulated speech) should perform almost equally well on the test sets of both origins. Additionally, any approach for noise reduction and robust feature extraction should improve the results for both sets by a similar magnitude. This is particularly interesting for robustness approaches based on the assumption of independence between speech and noise assuming an additive mathematical model like in Equation 4.2.

Evaluation Setup

Our evaluation is performed using cross validation for the six main speakers of each set, each speaker is used for testing in one iteration with all other speakers used for training. Each training set contains about 1400 utterances and each test set about 170 utterances. Even though the amount for training is very low, we can expect that the results give an indication, whether the above mentioned assumptions are correct. We further have two reference noisy speech sets without speaker overlap. That way we can evaluate both reference acoustic models on both test sets to identify, whether we have larger differences in the recognition accuracy of both reference models.

The speech recognition system used for evaluation is based on MFCCs. We include the first 12 coefficients plus energy and first and second order derivatives. Furthermore, we also test two common robustness approaches that were developed to deal (amongst others) with additive noise, namely spectral subtraction (compare Section 2.4.2) as well as the robust front end standardised by ETSI (compare Section 2.4.5).

Spectral subtraction is a basic idea forming a group of approaches. As an algorithm representing this group of noise reduction approaches we use the Matlab implementation *specsub.m v1.4*, which is part of the *voicebox*[6] toolbox. We use the algorithm with standard parameters plus Wiener filtering.

The ETSI robust front end is part of the ETSI standard ES 202 050 for robust distributed speech recognition. While the standard also includes several aspects of coding and decoding for distributed

[6] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

		clean	simulated	noisy
mfcc	clean am	51.01	25.96	30.71
	simulated am	37.50	38.16	35.22
	noisy am	32.85	30.60	43.06
ssub	clean am	49.79	27.93	32.00
	simulated am	42.87	34.90	34.09
	noisy am	39.62	29.13	39.14
etsi	clean am	49.01	34.53	35.61
	simulated am	41.91	38.67	36.83
	noisy am	38.97	33.72	42.63

Table 4.7: Phoneme accuracies for different setups of noisy speech. Clean speech, simulated noisy speech and realistic noisy speech test data is evaluated on the different acoustic models (am). Standard MFCCs and two robust front end processing methods, spectral subtraction (ssub) and the ETSI robust front end (etsi), are considered. The results for simulated noisy speech are not very similar to the results for the reference noisy speech.

systems, we only focus on the robust front end, which offers an advanced feature extraction framework for mismatch reduction. The basic steps performed by the front end are noise reduction based on a Wiener filter, SNR-dependent waveform processing of the noise reduced waveform, feature extraction of the cepstral coefficients and a blind equalisation of the coefficients.

For evaluation we train acoustic models for clean, noisy and simulated noisy speech for each feature extraction approach (standard MFCCs, MFCCs after Spectral Subtraction, ETSI features). For standard MFCCs and Spectral Subtraction we extract the cepstral coefficients with HTK based on the original speech signal and the speech signal after Spectral Subtraction. We use our standard feature set of 39 features as described in Section 3.3.6. In case of the ETSI processing, cepstral coefficients are directly provided by the front end. Monophone acoustic models for 44 phonemes of British English are trained on each set. 16 Gaussian mixtures are considered for each state of the HMM-based acoustic models.

As we have to deal with a rather small amount of acoustic training data, we evaluate using cross validation of the six major speakers of each set. In any case training and test data are disjoint in terms of data and speakers. Thus, all training and test set combinations provide speaker independent results.

Evaluation Results

Each test set (clean, noisy, simulated) is evaluated on all acoustic models of the same cross validation set. Table 4.7 shows the mean phoneme accuracies for all three data sets and the three approaches of feature extraction (MFCCs, Spectral Subtraction, ETSI).

The first lines show the results for standard MFCCs extracted by HTK. As expected the acoustic models in average perform best on the test set of the same acoustic attribute. Both reference noisy acoustic models trained on disjoint training sets perform almost equally well on both noisy test sets. Surprisingly, the acoustic models based on the simulated data generally show a comparatively low performance evaluated with simulated test data. On the other hand, the same acoustic models show a rather good performance when tested on data with different attributes (clean, noisy). This shows the ambivalent character of the simulated data which is basically clean speech but mixed with noise to simulate noisy speech. Still it is neither very close to the clean speech anymore nor much closer to the realistic noisy speech which it should simulate. This is opposed to the expectations when assuming that the simulated noisy data is very similar to the realistic noisy data. Furthermore, the low accuracy for simulated test data on the acoustic models of the same attribute and the comparatively high accuracy for the test data

from clean and noisy origin on the same acoustic models indicate a high variability of the simulated data. The simulated test data might also suffer from the lack of distinct peaks in the spectrum necessary to discriminate certain phonemes as discussed in Section 4.2.4. This assumption is further supported by the results of ETSI evaluation in the same table. Here, the performance of the simulated test data on the acoustic models of the same attribute benefits from the mismatch compensation of the front end, while both clean and noisy speech performance drop when using the robust ETSI features tested on the acoustic models of the same attribute (ETSI clean and ETSI noisy).

Both robustness algorithms to improve ASR when mismatch is present reveal interesting aspects. The ETSI robust front end generally shows good improvements whenever acoustic mismatch between models and test data is present, but slightly reduces the performance in case of no mismatch. Spectral Subtraction shows even better results than the ETSI front end for clean test data and acoustic mismatch (tested on noisy and simulated acoustic models). For simulated or noisy test data evaluated on the clean speech acoustic models the improvement is much less significant. Furthermore, the results suffer in case of no mismatch or testing simulated data on noisy models and vice versa. These differences are quite surprising. A probable explanation is the higher variability covered by simulated and noisy data. Considering 16 Gaussian mixtures, a higher variety of possible distortion caused by Spectral Subtraction might be covered by the related acoustic models compared to the clean speech models with Spectral Subtraction. More detailed evaluations with a larger amount of data might help to provide an answer to these particular results.

4.2.6 Conclusion

Many approaches to robust ASR are evaluated on simulated noisy speech data assuming that this data is very similar to realistic data, and thus, the results are valid in real-life applications as well. In our evaluation we could show that this assumption is not necessarily true. Even if we are able to simulate noisy speech from clean speech and noise samples which are very similar concerning speech and noise characteristics compared to a reference set, the performance is not comparable at all. This evaluation is even based on rather ideal data with identical channel characteristics for all data from clean speech, noisy speech towards noise samples. In other setups of noise simulation this is usually not the case.

In general this evaluation shows the complexity of additive noise, which is opposed to the assumption not just additive. In [64, 65] the Lombard effect is investigated that changes the way a speaker is speaking in a noisy environment. This effect breaks with the assumption that speech and noise are independent and can be modelled purely additively. Thus, we assume that the major difference between simulated noisy speech and noisy speech leading to the difference in the ASR performance are caused by the Lombard effect, which is only present in the realistic noisy speech. For a better simulation and better matching results, an improved noise model might be necessary reflecting the effects of Lombard speech. Unfortunately, this effect is highly non-linear and dynamic ([64]) and thus difficult to model. The difference in simulation also indicates that robustness approaches considering additive noise for noise removal will only remove part of the distortion in the best case.

In any case, common speech features show already a high variability in the cepstral coefficients in similar acoustic conditions as we could see on four examples. This variability of the features further increases with varying acoustic conditions. For exactly the same utterance but with artificially added noise the cepstral values already show significant differences. While certain normalisation techniques (compare Section 2.4.1) are able to cope with some of the variations, such techniques like global cepstral mean or cepstral gain normalisation, for example, still just correct variations in the trend of the features but will not suffice in terms of a proper compensation of many other short time effects and variations. As our experiments indicate even “simple” additive noise seems to have rather complex influences on

the speech features.

4.3 Microphone Channel Effects

In the previous section we analysed the effect of additive noise on the performance of ASR and showed the insufficiency of the commonly assumed mathematical model of additive noise. One of the reasons for this insufficiency are non-linear distortions influenced by environmental noise due to the Lombard effect. Another very important source for non-additive acoustic distortion is the effect of the channel. This includes all effects caused by hardware, filtering, coding, etc. while recording and transmitting a speech signal. In this section we investigate the influence of microphone channel distortion by comparing two sets of synchronous recordings of various utterances captured with two very different microphone technologies. One of the microphones is a high quality close-talk microphone, while the other one is a throat microphone picking up vibrations directly from the larynx. Both recordings only differ in the channel effects and — in case of background noise — in the level of recorded additive noise. As we can compare exactly the same utterances, differences in speaker characteristics and speaking style are not present at all. A general evaluation of the effect of the microphone channel on ASR was already presented in the beginning of this chapter in Section 4.1. Here, we will dig deeper to show channel differences and their influence on ASR. Major aspects of the following evaluation of the microphone channel effects we also published in [4].

4.3.1 Related Work

Throat microphone speech and combinations of close-talk and throat microphones for ASR have been researched and evaluated in several publications. In [124, 125] a throat microphone and a close-talk microphone are combined and used for robust speech recognition in noisy environments. The authors in [124] use a probabilistic optimum filter (POF) mapping algorithm to map the noisy mel-cepstral features extracted from both microphone signals to clean speech features. These features can then be used by a standard speech recogniser trained on large amounts of clean speech. Significant improvements compared to a POF mapping of the single close-talk channel are reported. A detailed analysis on both microphone signals is not provided.

In [125] a detailed description of alternative acoustic sensors with a focus on the throat microphone is given. Furthermore, a brief quality comparison between throat microphone and close-talk microphone for ASR are given, mainly stating that unvoiced sounds are attenuated by the throat microphone while nasals or voiced plosives, for example, are relatively amplified. In a final evaluation on a hybrid HMM/ANN (Artificial Neural Network) speech recognition system, a combined approach with voice activity detection from the microphone signal and a combination of the emission probabilities of both signals show good improvements compared to a single close-talk and a single throat microphone recognition.

The authors in [126] aim at an adaptation of an existing system for soft whisper recognition using a throat microphone. They compare several steps of filtering and MLLR adaptation in their work and further give a brief hint on some of the differences between standard and throat microphones.

The most detailed analysis on the characteristics of throat microphones compared to standard close-talk microphones is given for Hindi in [127]. Comparing the differences of the spectra for different vowels and analysing various more general characteristics of sound units recorded by both microphones, they motivate a mapping approach using a neural network to enable a mapping of the speech spectra of the throat microphone signal to speech spectra of close-talk microphones to improve the speech quality

of throat microphone signals.

We want to confirm and extend several of the findings of the related work by a more detailed evaluation on both microphone signals with a focus on aspects relevant for ASR.

4.3.2 Microphone Channels

The speech data for evaluation is part of the MoveOn Corpus described in Section 3.3. We were able to record speech with two very different microphone technologies synchronously enabling a direct comparison of both recorded signals without influences of speaker and speech variabilities. The two microphone technologies compared here are a high-quality close-talk microphone and a noise robust throat microphone. Both microphone channels are briefly described in the following part of this section.

Close-Talk Microphone

Standard close-talk microphones are well-known to most people. Different transducer concepts as well as different characteristics, considering polar pattern, frequency response and dynamic range, exist. But the fundamental concept is to record airborne sound. In this work a close-talk microphone (AKG C417 - condenser lavalier microphone) with an almost linear frequency response in the spectral range of speech (up to about 8 kHz) is used. The microphone is capable of recording even high sound pressure levels with low distortion. The polar pattern is omnidirectional. The microphone was placed within the motorcycle helmet about 4 cm right of the centre of the speaker's mouth providing near field speech recordings.

Throat Microphone

Throat microphones are more robust to environmental noise than usual close-talk microphones and, hence, are often used for communication in noisy environments in military and other applications. A throat microphone is put around the neck with the transducer placed on the larynx with slight pressure. Instead of airborne sound solid-born sound is picked up directly from the larynx. Thus, a throat microphone is less prone to any environmental noise, but does not provide a good frequency response in the range of speech. A standard single transducer throat microphone — the Alan AE 38 Throat Microphone — provides an alternative speech signal for evaluation.

4.3.3 Influences on Speech Characteristics

Close-talk and throat microphones have substantially different qualities regarding spectral characteristics and signal-to-noise ratio (SNR). The SNRs and some spectral characteristics of the microphones are compared in this section.

Signal-to-Noise Ratio

The SNR for both microphone signals is estimated by NIST STNR of the NIST Quality Assurance Package [7]. In Figure 4.8 the average SNR of the throat microphone signal is plotted versus the related SNR of the close-talk microphone signal for synchronously recorded utterances. Additionally, a histogram of the SNR values of the close-talk signal is shown.

For low SNRs the throat microphone is generally less affected by environmental noise, i.e., the average SNR is higher than the SNR for the close-talk microphone signal. For less noisy signals (20dB

[7] NIST SPQA 2.3, `http://www.itl.nist.gov/iad/mig/tools/`

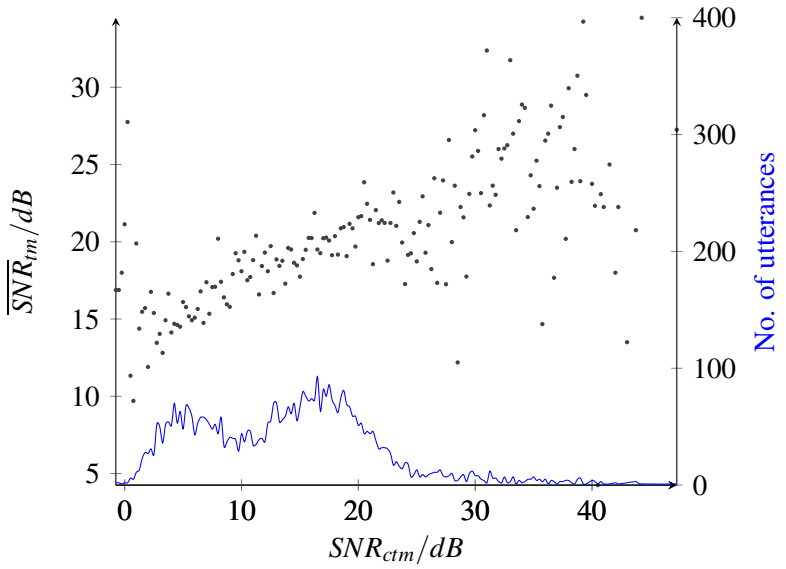


Figure 4.8: Comparison of SNRs for close-talk microphone and throat microphone signals. The blue line shows the distribution of the estimated SNR_{ctm} of the close talk microphone for all utterances. The relation between estimated SNR_{ctm} and average estimated $\overline{SNR}_{tm}$ of the throat microphone is indicated by the dots and shows a linear tendency but with rather high variances in case of a low number of sample utterances available for averaging.

and higher) the close-talk signal usually provides a clearer speech signal with a better frequency response resulting in higher SNR values compared to the throat microphone signal. A linear dependency between both SNRs can be approximated from Figure 4.8. The variance of the SNR values along the approximated line is rather high, which is noticeable for high and low SNRs where the variances were not averaged out due to a low number of measured values (see histogram in the same figure). The high variance might have two explanations, first, the estimation with the NIST tool is not very precise, and second, a dependency between the signal-to-noise ratio of both signals might exist but might also be weak.

Speech Characteristics

Close-talk microphone signals sound more natural compared to signals from a throat microphone. The latter resemble sometimes “metallic” sounds. This is due to limitations in the frequency range of the signal. We will analyse the speech characteristics for both signals in the following paragraphs and show typically differences on several examples of close-talk and throat microphone speech.

While spectrograms of close-talk and throat microphone signals are quite similar in the lower frequency range, the high frequencies differ substantially. We can see this effect, when comparing the spectrograms (a) and (c) as well as (b) and (d) of exactly the same utterance synchronously recorded by both microphones in Figure 4.10 (and also in Figure A.2). Also depending on the phoneme, frequencies from 4 kHz and above appear to be less present or entirely inaudible in the throat microphone signal. This limitation is due to the type of sensor and the location of the transducer relative to the speech source. Apparently, the signal conducted by skin near the larynx (as in the case of a throat microphone) is of limited bandwidth, which is caused by the low transducer sensitivity for these frequencies and by the resilience of the skin attenuating mainly higher frequencies.

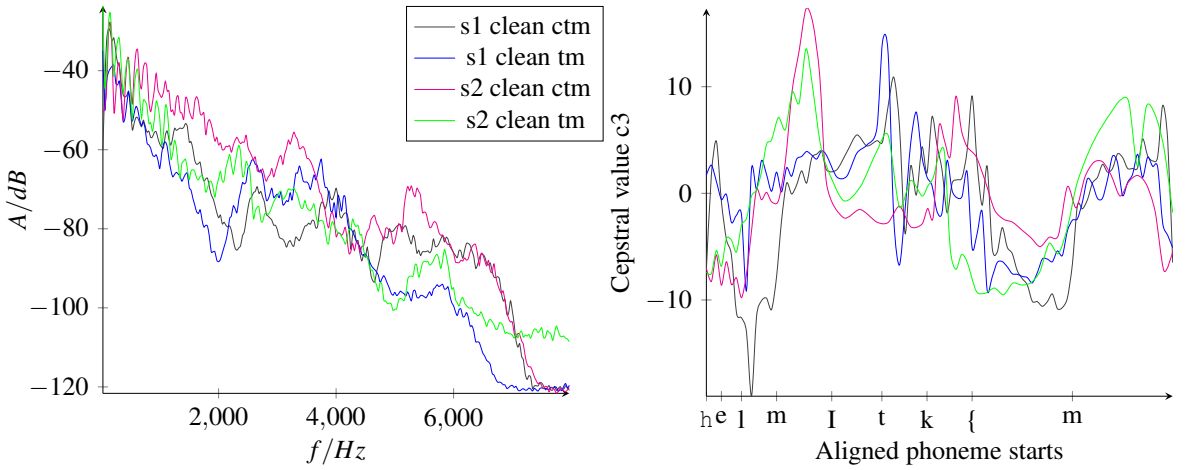


Figure 4.9: Example for influences on spectrum and cepstral values caused by speaker and channel. The figure on the left-hand side shows the spectrum for the utterance "Helmet Cam" for the examples from Figure 4.10 of two different speakers ($s1$ and $s2$) and throat (tm) and close-talk microphone (ctm). The figure on the right-hand side compares the aligned third cepstral coefficients for the same examples.

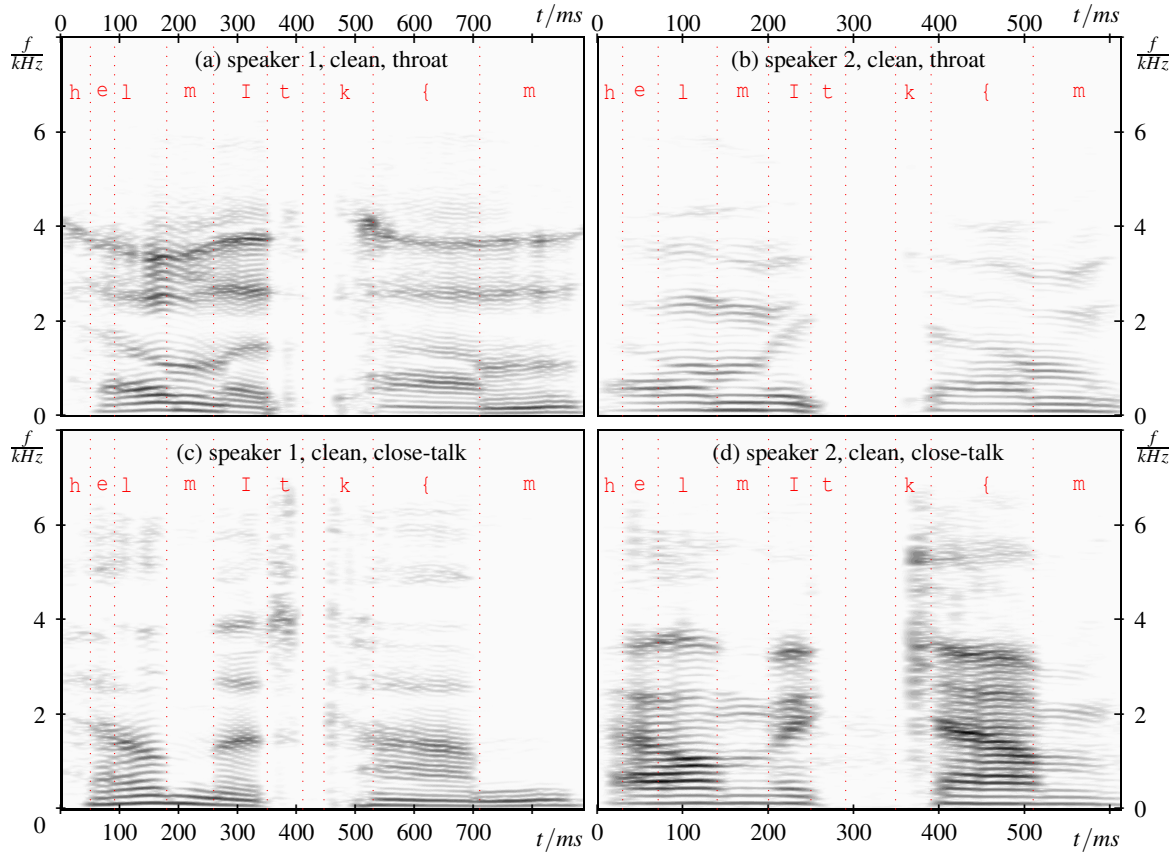


Figure 4.10: Spectrograms for channel and speaker variations for two speakers. The spectrograms show the utterance "Helmet Cam". (a) and (c) as well as (b) and (d) are synchronous recordings of the throat and close-talk microphone of the same utterance of the same speaker $s1$ or $s2$. High frequencies are practically missing in the throat microphone signal. On the other hand, certain phonemes (especially the /m/) seem to be better represented in the throat microphone signal.

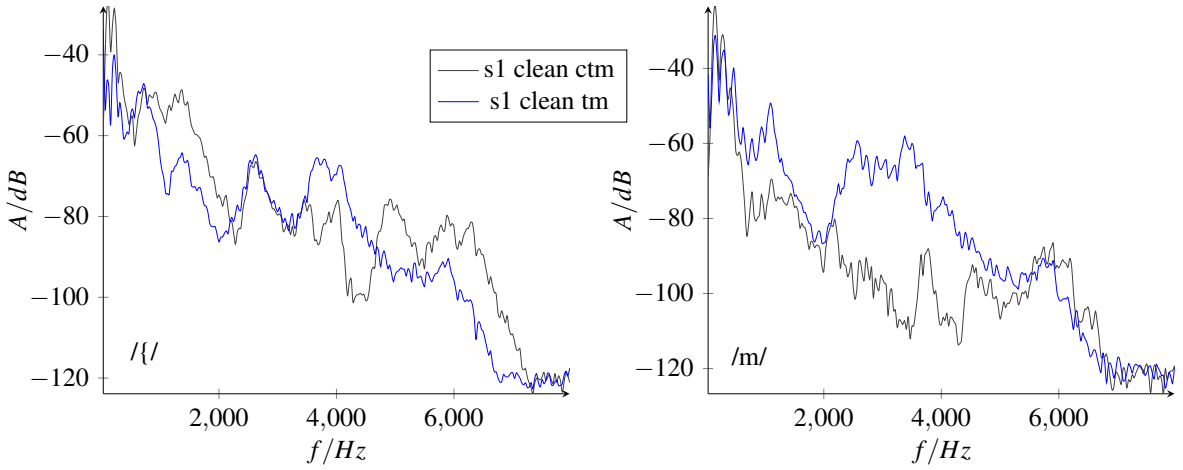


Figure 4.11: Influence of the channels on the spectrum of the phonemes /{/ and /m/. The spectra for both phonemes show various differences for throat microphone (*tm*) and close-talk microphone (*ctm*) channel for synchronous recordings of speaker *s*1. Differences include frequency dependent differences in the peaks' amplitudes as well as peaks missing in one of the signals while being clearly visible in the other one.

Generally, we can expect that many differences are closely related to the production of a phoneme and where in the speech production system a certain phoneme is mainly formed, as throat microphones pick up the vibrations close to the vocal cords and close-talk microphones on the other end of the speech production close to the lips. This explains, for example, why the phoneme /*m*/ for both occurrences in the utterance "Helmet Cam" has a more dominant representation in the throat microphone signal (*a*) and (*c*) than in the close-talk microphone signal (*b*) and (*c*) in the spectrograms in Figures A.2 and 4.10. This can also be seen in Figures 4.11 and 4.12, left-hand side, where the energy in the throat microphone signal compared to the close-talk microphone signal is much higher for most frequencies. This is not the case for the phoneme /{/ in the same figures, left-hand side. Interestingly, the energy for both phonemes of the throat-microphone signal is rather similar for both speakers while the energy of the close-talk microphone shows a much higher difference between the speakers. For both speakers and phonemes we can see that many of the characteristic spectral peaks are common for both microphone channels, while the peaks and the general curve for other frequency bands (e.g. between 4 kHz and 5 kHz in both figures) show major differences.

The aligned third cepstral coefficient[8] for the complete utterance "Helmet Cam" shows a similar tendency for both speakers and microphone channels, but also rather large differences between the two speakers, especially in the peaks for the phoneme /*m*/ and the phoneme /*t*/. But also for the two microphone channels the cepstral values show rather large differences in certain parts of the utterance with a difference of more than 10 units (out of a maximum range of about 30 units) in some cases like for phonemes /*e*/ and /*l*/ for speaker 1 and phonemes /*k*/ and /{/ for speaker 2.

Jou et al. [126] showed that a throat microphone signal cannot be modelled by simply applying a (sigmoidal) low pass filter as spectral differences between close-talk microphone and throat microphone are dependent on the phoneme. This dependency on the phoneme is caused by the different parts of the vocal tract that are responsible for the articulation of each phoneme, and the position and technique of the two microphones which unequally influence their sensitivity to sounds caused by the different parts

[8] The third coefficient is arbitrarily selected to demonstrate the influence of the microphone channel on the cepstral coefficients. While the particular influence on each coefficient is usually different, one will find many similarities between the cepstral features in the general pattern and magnitude of changes.

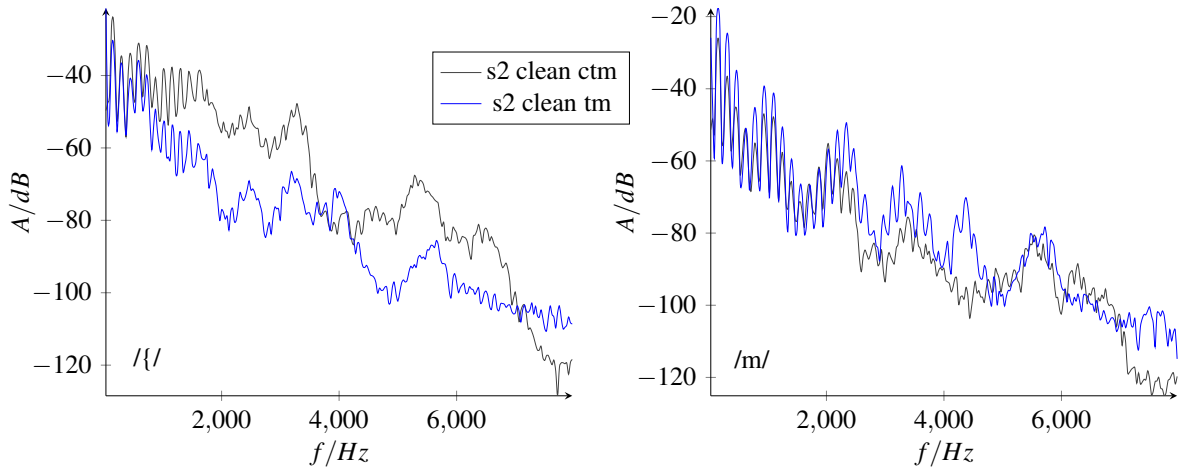


Figure 4.12: Influence of the channels on the spectrum of the phonemes /{/ and /m/. The spectra for both phonemes show various differences for throat microphone (tm) and close-talk microphone (ctm) channel for synchronous recordings of speaker $s2$. Several effects observed for speaker $s1$ in Figure 4.11 are also visible here. Magnitudes of differences vary between both speakers and the frequency characteristics of throat and close-talk microphone channels for phoneme /m/ are much more similar for speaker $s2$ compared to speaker $s1$.

of the vocal tract. They further claimed that consonants such as $/m/$ and $/s/$ differ significantly in both signals: The $/m/$ in a throat speech spectrum looks more like a vowel, while $/s/$, with higher energy at high frequencies and lower energy at low frequencies, is hard to hear in a throat microphone signal and subsequently also hard to recognise. In case of the $/m/$ we can also see this effect in Figures 4.11 and 4.12.

In general we can assume that certain phonemes are better represented in the throat microphone signal while others are better represented in the close-talk microphone signal. Shahina and Yegnanarayana [127], for example, stated that voiced stop consonants like $/d/$ and $/g/$ are represented better in case of throat microphone speech. Another candidate for a better representation in the microphone channel seems to be the $/m/$ with its more harmonic, vowel-like structure. The examples mentioned above indicate that ASR on the throat microphone signal compared to the close-talk microphone signal will perform better on certain types of phonemes with harmonic and voiced characteristics and worse on other phonemes with more unvoiced characteristics and their energy mainly in higher frequency bands. In the following evaluations of the ASR performance on both microphone channels we will analyse these aspects in more detail.

4.3.4 Evaluation of ASR Performance

In the previous section, several differences between the signals of a close-talk and a throat microphone were stressed. Here, the influence of the different characteristics on the speech recognition performance is evaluated.

Evaluation Setup

The training and evaluation of the acoustic models is performed by HTK. The setup is identical to the preliminary evaluations in Section 4.1.3, but we do not consider any robust front end processing here. Furthermore, we also evaluate close-talk and throat microphone acoustic models on both clean and noisy test set. In short, we prepare two sets of acoustic models for each microphone channel based on the clean

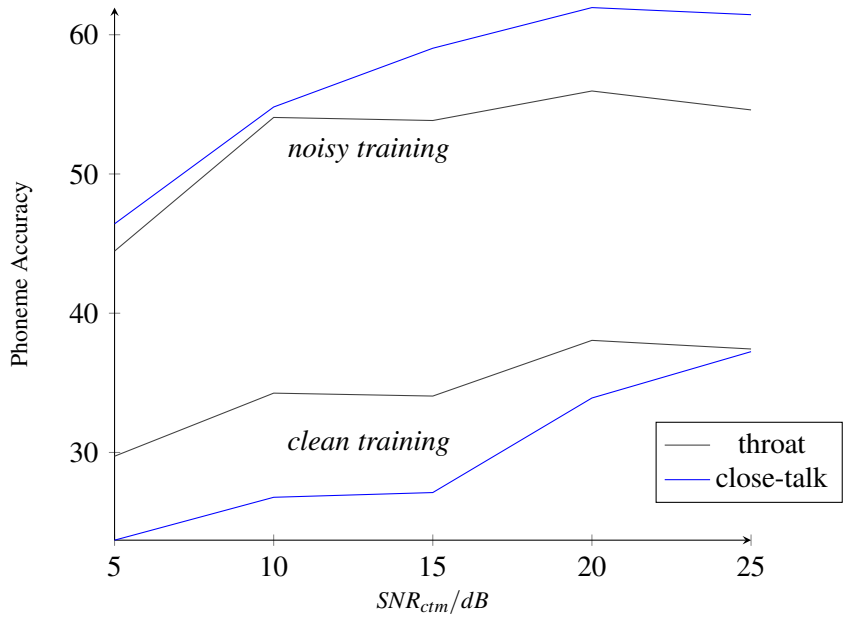


Figure 4.13: ASR performance of close-talk and throat microphone. The figure (extracted from Figure 4.3) demonstrates the influence of matched (noisy training) and mismatched acoustic conditions (clean training) on the recognition performance. The high quality recordings of the close-talk microphone (*ctm*) provide better results than the throat microphone (*tm*) in matched conditions, while the throat microphone is less affected by the acoustic environment (compare results for low SNRs).

and the noisy training sets with the amount of data stated in Table 4.2. The acoustic models are based on 12 static MFCCs with energy and first and second order derivatives. Cepstral mean normalisation is performed. Due to the small amount of data, monophone acoustic models are trained. Speaker and environmental variabilities are taken into account by using 16 Gaussian mixtures per state.

Phoneme recognition without high-level syntactic relations is performed assuming an equal distribution of all phonemes to evaluate the acoustic effects. Two test sets are used, a clean speech test set with 417 clean utterances of two different speakers and a noisy speech test set of 588 noisy utterances recorded from three speakers while riding a motorcycle. Both test sets do not overlap with the training set.

Evaluation Results

The results of Table 4.6 for noisy test data evaluated on different sets of acoustic models and front ends are pictured. Throat microphone data and ETSI front end trained on mismatched clean conditions improve results compared mismatched close-talk acoustic models. Non-mismatched conditions of noisy close-talk acoustic models perform best.

Figure 4.13 compares the phoneme accuracies for both microphone channels and different SNRs. Only the noisy test data is used, grouped in 5dB steps by estimated SNR of the close-talk channel. These results were already presented and compared to the ETSI front end performance in Section 4.1.3. The mismatch between clean training data and noisy test data is lower for the throat microphone signal than for the close-talk microphone signal as environmental noise has less impact on the throat microphone. This was already indicate by Figure 4.8 showing higher estimated SNRs for this microphone compared to the close-talk microphone for noisy utterances below an SNR of 20dB in the close-talk microphone

signal. Thus, the results of the recognition at low SNRs are distinctively better for the throat microphone in case of clean training data. For acoustic models based on noisy data (without relevant mismatch), the close-talk signal outperforms the throat signal — especially for high SNRs — due to the better frequency characteristics of this microphone discussed in Section 4.3.3.

In the same section we also discussed that the performance for certain phonemes should still be better for the throat microphone channel due to the microphone position and speech production process. Thus, we also determine the phoneme accuracy differences $\Delta a(p)$ for each phoneme p between close-talk microphone (ctm) and throat microphone (tm):

$$\Delta a(p) = a_{tm}(p) - a_{ctm}(p) \tag{4.6}$$

We further determine the influence of the three types of errors (number of substitutions S, insertions I and deletions D — compare Section 2.5.1) causing the phoneme accuracy differences. The influence of these types of errors on the difference in phoneme accuracy rate is as follows replacing the accuracy in Equation 4.6 with the phoneme error rate PER based on Equation 2.52:

$$\Delta a = -PER_{tm} + PER_{ctm} = -\Delta s - \Delta d - \Delta i \tag{4.7}$$

Differences in substitution, deletion and insertion are described in this equation by their difference rates Δs, Δd and Δi defined as follows (with S, I, and D as described above and the number C of correctly recognised phonemes):

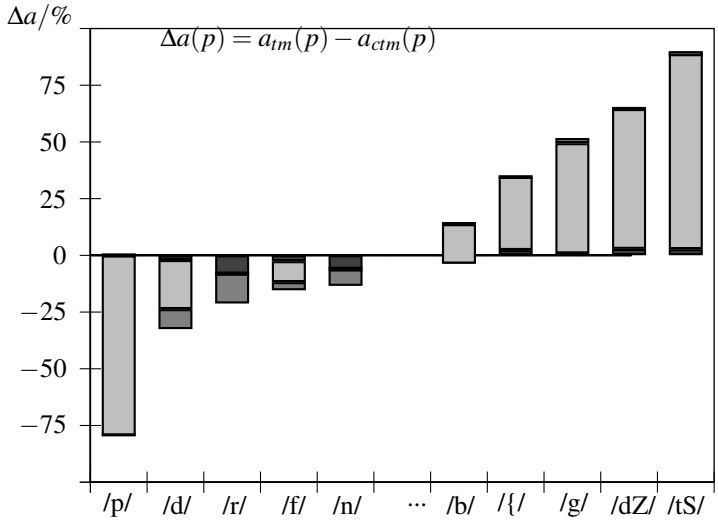
$$\Delta s = \frac{S_{tm} - S_{ctm}}{S + D + C}, \tag{4.8}$$

$$\Delta d = \frac{D_{tm} - D_{ctm}}{S + D + C}, \tag{4.9}$$

$$\Delta i = \frac{I_{tm} - I_{ctm}}{S + D + C}. \tag{4.10}$$

The bar chart in Figure 4.14 shows the absolute differences of phoneme accuracies of those phonemes with largest differences between throat microphone and close-talk microphone. Negative values indicate lower accuracies for the throat microphone. The figures are estimated with clean speech models and clean speech test data only to exclude effects caused by background noise. Influences of phoneme correctness, insertion and deletion contributing to the overall phoneme accuracy are shown in different grey scales. Especially $/p/$ is recognised with a significantly higher accuracy by the close-talk microphone mainly due to fewer insertions of this phoneme. A probable reason might be a similarity to non-speech sounds like gulps or other reflexes close to the larynx. The lack of typical characteristics of the plosive $/p/$ in the higher frequency bands in case of the throat microphone makes the recognition of this phoneme also more difficult. The phonemes $/tS/$, $/dZ/$ and $/g/$, on the other hand, are recognised better by the throat microphone. The difference is again mainly due to a difference in the number of insertions while phoneme correctness and number of deletions are almost equal. The phonemes $/tS/$ and $/dZ/$ are similar to breathing and other non-speech airborne sounds, which might cause additional insertions when captured by the close-talk microphone. The phoneme $/g/$ is assumed to be recognised better from the throat microphone signal as we already discussed in Section 4.3.3.

As phonemes can further be grouped by similarities in their production, we also compare the accuracy rates for different typical phoneme groups in Table 4.8. We roughly grouped all phonemes into five groups plus silence (for monophones $/sil/$ and $/sp/$) depending on their particular articulation. Instead of a phoneme recognition we perform phoneme group recognition modelling each group of phonemes



dark gray: $-\Delta s$, gray: $-\Delta d$, light gray: $-\Delta i$ (Equations 4.8-4.10)

Figure 4.14: Phoneme accuracy differences Δa for throat to close-talk microphone. The difference in phoneme accuracy Δa between throat *tm* to close-talk microphone *ctm* indicates, which phonemes are better recognised using a certain microphone. Phonemes to the left-hand side of the chart show a higher phoneme accuracy rate when using the close-talk microphone, phonemes to the right-hand side of the chart show a higher phoneme accuracy rate when using the throat microphone.

	ctm		**tm**	
	clean	**noisy**	**clean**	**noisy**
nasal	93.59	94.35	91.89	96.35
plosive	97.09	96.96	95.31	95.78
vowel	96.32	97.00	90.83	96.78
liquid	87.74	86.09	72.03	77.78
fricative	96.30	89.24	94.71	95.10
sil	92.64	91.81	97.17	98.21

Table 4.8: Specific phoneme group accuracy rates for close-talk and throat microphone. The accuracy rates for phoneme group recognition for close-talk (*ctm*) and throat microphone (*tm*) indicate advantages considering silence detection and fricative recognition for the throat microphone in particular in noisy environments. Liquids are better recognised using a close-talk microphone.

instead of monophones. Both test sets (clean and noisy) are evaluated using the appropriate clean or noisy speech models for each set.

Vowels and other voiced phonemes are produced by vibrating the vocal cords in the larynx resulting in an excitation signal of a fundamental frequency in the lower frequency band (up to 300 Hz). The vocal tract cavities modify this signal causing multiple resonance frequencies which are referred to as formants [128]. These phonemes (nasals and vowels) can be recognised fairly well using both kinds of microphones. The only exception are the clean office recordings of the throat microphone as we can see in our results. The drop in performance for these recordings might be caused by the female speakers with attenuated speech signals caused by low pressure of the throat microphone transducer against the larynx (compare Section 4.1.3). In case of unvoiced consonants (unvoiced fricatives and plosives like $/f/$ and $/t/$) there is no fundamental frequency and the sound contains considerably less energy which is further concentrated in higher frequencies. This probably makes it difficult to recognise this kind of sounds using a throat microphone.

Liquids, which are voiced consonants without friction like $/l/$, $/r/$ or $/w/$, are generally recognised worse than other phonemes — especially using the throat microphone signal for recognition. For the close-talk microphone the accuracy rate of the fricatives drops significantly for noisy data. Fricatives usually suffer most from environmental noise due to their noise like characteristics. Thus, this problem does not occur for the throat microphone signal as it is less prone to environmental noise. For similar reasons silence can be recognised significantly better from throat microphone signals. Generally, phoneme group recognition on the throat microphone signal performs better for noisy speech than for clean speech. This can be explained by the larger number of utterances in the noisy training set and the attenuated signal for the female speakers in the clean throat microphone sets.

4.3.5 Conclusion

The results of the evaluation show both similarities and major differences between throat and close-talk microphones for ASR. The differences are highly variable, dependent on the frequency and the phoneme with a tendency of a low-pass filtering for the throat microphone. These high variability makes it very difficult to transform the signal or the features of a throat microphone channel into a signal or into features similar to the ones of a close-talk microphone channel and vice versa.

The ASR performance of the throat microphone channel in noisy conditions is mainly influenced by the robustness towards environmental noise, on the one hand, and the different frequency characteristics, on the other hand. While the throat microphone signal is less influenced by environmental noise and, hence, performs well even in rather noisy conditions, it lacks a good frequency response compared to close-talk microphones. Thus, without any mismatching environmental conditions the close-talk microphone usually performs better due to the better quality of the speech signal. While the throat microphone enables improved recognition results for some specific phonemes (especially nasals and silence), recognition accuracy based on the throat microphone signal especially decreases for plosives and liquids.

All in all, the high quality close-talk microphone performs slightly better than the throat microphone due to the better frequency response. In [69] several approaches trying to overcome the disadvantages and problems of the throat microphone channel are evaluated. The results in that work with no significant improvements indicate that channel and speaker induced effects seem to be rather complex and difficult to compensate with a general linear approach or by alternative features as often suggested for throat microphone ASR.

4.4 Hardware, Transmission and Coding Effects

In Section 4.3 we discussed the influence of the microphone channel on the speech signal and the speech recognition performance. In addition to the microphone channel various other channel effects, for example, of other hardware components or the transmission channel including coding and decoding effects can further influence the speech signal (Section 2.3). We avoided these effects for the MoveOn Corpus by using the same hardware setup without wireless transmission and without coding and decoding of the signal.

In this section, on the other hand, we want to evaluate exactly these hardware and transmission effects. As the MoveOn Corpus is insufficient for this purpose as we avoided such effects, we created another corpus for evaluation simulating step by step the different channel and coding influences caused by a TETRA radio channel. We decided to base this evaluation set on our medium to large sized speech corpus of clean broadcast data used for training and testing of our LVCSR system. The creation of this evaluation corpus was already discussed in Section 3.2.

In the following, we report on several experiments to identify the challenges encountered by ASR in case of a low bandwidth transmission channel severely influencing the speech signal. We further separate the various aspects of the TETRA transmission channel to understand, which parts are most critical for the speech signal and the ASR performance. This evaluation shows the variability of such channels and the caused effects. Parts of our work on the TETRA radio data was published in [6, 7].

4.4.1 Related Work

Terrestrial Trunked Radio (TETRA) [104] is a standard for digital trunked radio systems, first published by the European Telecommunications Standards Institute (ETSI) in 1995. It has been designed for robust speech transmission and indeed is used for public safety networks across Europe, Asia and other countries. However, its influence on automatic speech recognition (ASR) has rarely been analysed.

The objective speech quality of the TETRA codec is examined in [129]. While focussing more on the speech quality degradation in correlation with the bit error rate, the findings are, as the authors mention themselves, somewhat inconclusive. [130] offers an extensive overview of the TETRA encoding/decoding performance, and on package delay and throughput in an overall architecture, with special focus on transmission errors and co-channel interference.

Scientific papers analysing the TETRA encoding impact on natural language processing by automatic means are scarce. The authors of [131] analyse the TETRA codec on the speaker recognition performance. They do not only work on the audio signal, but also make direct use of the linear prediction coefficients that are computed by the TETRA encoder (Section 3.2.2). Simply taking the decoded speech signal performs worst and seems to be the hardest setting. [73] is one of the few papers employing actual TETRA data in their recognition setup. On a small corpus of spoken German digits, they show that the TETRA codec performs poorly in comparison to the plain signal, to a 16 kbit/s Code-Excited Linear Prediction (CELP), and to a GSM codec.

In related work the rather complex characteristics and poor speech quality of the TETRA channel caused by coding/decoding effects, transmission errors, co-channel interferences etc. are already stressed. In the following we will separate some of these effects and focus our evaluation on a step by step analysis of the degradation of the signal in the context of ASR on the TETRA channel.

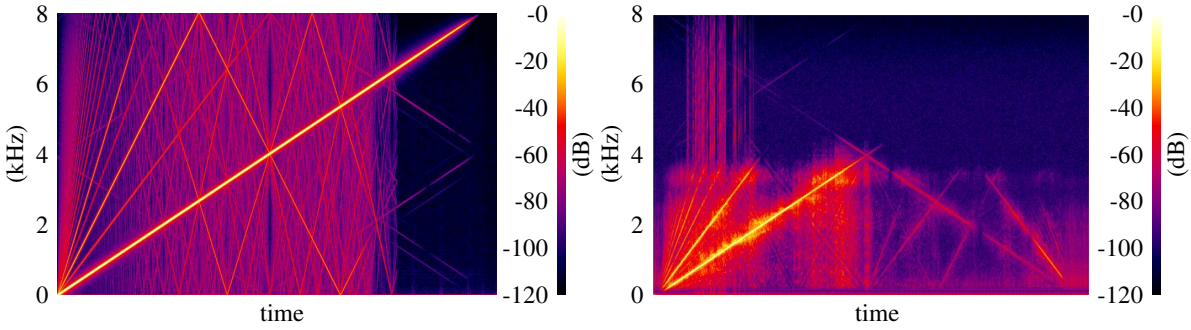


Figure 4.15: Spectrogram of the test signal from closed-loop and received over TETRA. A sweep from 0 to 8 kHz is recorded from a closed-loop connecting the line output with the line input of a high quality sound card (left-hand side) and when transmitted via TETRA (right-hand side). Harmonic distortion occurs in both examples. The TETRA signal further introduces additional noise and low-pass filtering effects cutting frequencies beyond 4 kHz.

4.4.2 Preliminary Evaluation

In a preliminary evaluation we transmitted a synthesised frequency sweep from 0 to 8 kHz via the actual TETRA channel using the hardware described in Section 3.2.2 to characterise the frequency characteristics of the actual TETRA channel. In the setup, the TETRA radio station is used as sender, having the input signal fed in via the headset connector. The hand-held device acts as the receiver, and the signal is recorded from the line output. Figure 4.15, right-hand side, illustrates the drastic quality deterioration. The encoding and subsequent transmission adds noise to the whole spectrum and also suppresses all frequencies above 4 kHz. Furthermore, significant harmonic distortion is visible and audible in the recorded sweep signal. As harmonic distortion is a typical distortion introduced by many hardware components (compare Section 2.3.3), we conducted a separate experiment where we re-recorded the signal without real TETRA transmission. We played back the sweep and fed the line output of a high quality sound card to the line input of the sound card and recorded the incoming signal ("closed-loop"). This is practically the same setup as we used to transmit the signal via TETRA but omitting the TETRA hardware this time. The resulting spectrogram is shown in Figure 4.15, left-hand side. Thus, we could generally attribute the massive amounts of harmonic distortion as witnessed in the spectrogram to the audio hardware. In Figure 4.16 we see the spectrum of the original sweep compared to the closed-loop setup and the actual TETRA transmission. The closed-loop signal is almost identical to the original signal except for a gain of about 1.5dB probably caused by the introduced harmonics. An actual TETRA transmission adds additional signal distortion and particularly affects the frequencies below 300 Hz and above 3400 Hz, which are considered to be less important for speech intelligibility. But also frequencies between 2400 Hz and 3400 Hz already suffer from relatively high attenuations.

While we could show that the harmonic distortion is probably originated by the audio hardware in the TETRA radio equipment, which usually causes harmonic distortion (also compare THD in Section 2.3.3), we further assume that the design of the adaptive code book in the TETRA codec emphasises them. The reason might well be that, since the codec is optimised on human speech intelligibility, special focus is on the preservation of harmonics produced in voiced phonemes, at the cost of further amplified harmonic distortions in the signal. We will analyse this aspect later in this section.

4.4.3 Influences on Speech Characteristics

The influence of the different steps of the TETRA channel and the real radio signal (TETRA Radio) is visualised in the spectra in Figure 4.17. The spectra are derived from the same utterance ("Tagesthemen") for all examples. The peaks and trends for all signals are quite similar and except for the TETRA

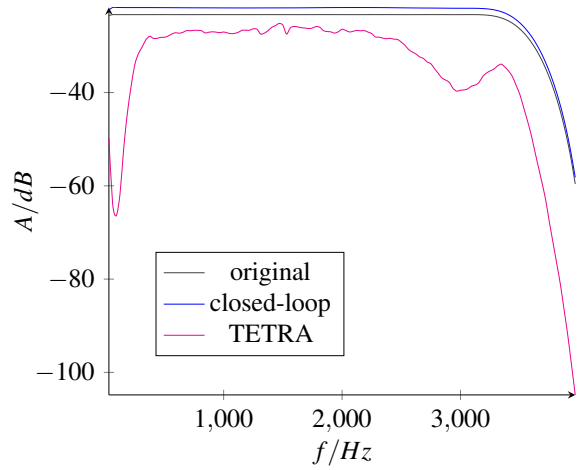


Figure 4.16: Frequency analysis of the sweep for closed-loop and TETRA transmission. The curve of the spectrum of a sweep recorded from a closed-loop is very similar to the original signal's spectrum with an increased average amplitude for all frequencies due to introduced harmonics. The spectrum of the TETRA transmitted signal shows a general attenuation of all frequencies and in particular for frequencies below about 300 Hz and above 2.5 kHz.

Radio channel are almost identical for the lower frequencies. The TETRA Radio shows the most significant differences when compared to the other signals. One reason are the filtering effects of the real TETRA channel visible in the sweep analysis (Figure 4.16) affecting the frequencies of the utterance in Figure 4.17, left-hand side, accordingly: Low frequencies below 300 Hz are practically filtered out completely, while frequencies above 2400 Hz are significantly attenuated.

In Figure 4.17, right-hand side, we see the third cepstral coefficient for the 8 kHz clean speech, the TETRA codec and the TETRA radio signal. While the coefficients for clean speech and TETRA codec are rather similar, the same coefficient for the TETRA radio signal shows two major differences. As we have some additional tenth of seconds before and after the speech signal to avoid cutting the speech while recording, feature extraction and cepstral normalisation is performed on a larger set of cepstral feature vectors. Thus, the features are not exactly centred for the aligned frames seen in Figure 4.17. Furthermore, some additional distortion of the TETRA radio coefficient is obvious, especially between frame 40 and 60, which we assume is caused by the harmonic distortion.

4.4.4 Evaluation of ASR Performance

In the following experiments we separate the different stages of a TETRA transmission and analyse the influence of each stage on the speech recognition performance. Furthermore, we will compare the results of each step of the simulated TETRA channel with the signal actually transmitted via the TETRA radio and discuss the degradation of the recognition results.

Evaluation Setup

As opposed to the experiments before, the TETRA Corpus is based on our speech corpus designed for large vocabulary continuous speech recognition (LVCSR). Thus, we decide to use an LVCSR system for evaluation incorporating triphone acoustic models and a complex language model. We employ HTK for feature extraction using our standard set of 39 features (Section 3.3.6).

Due to our decision for using a LVCSR approach we require a language model. We make use of the

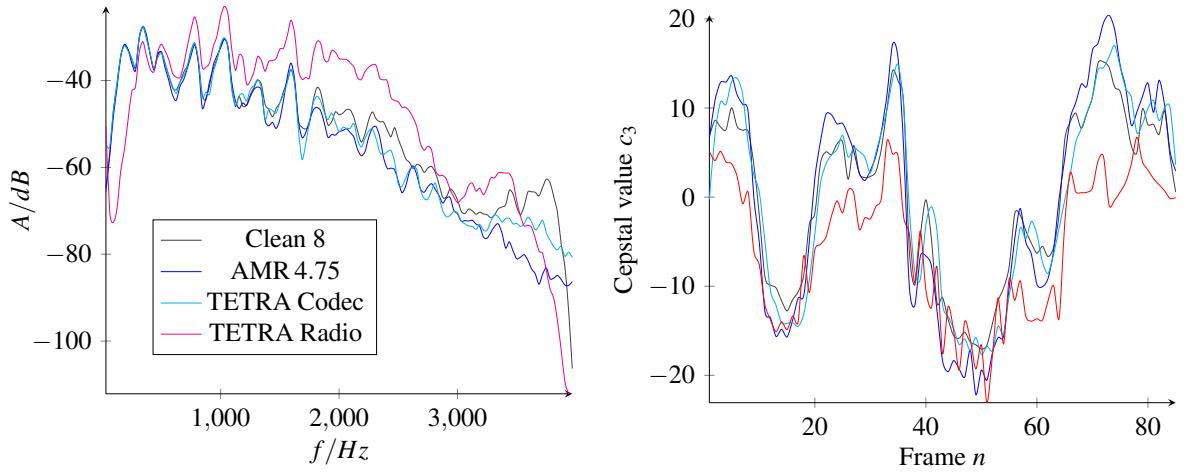


Figure 4.17: Influence of the TETRA channel on the frequencies and cepstral values. The influence of the AMR 4.75 and TETRA coding as well as the influence of the actual TETRA transmission of the original clean speech utterance (Clean 8) is presented for the spectrum (left-hand side) and the cepstral coefficients (right-hand side). AMR and TETRA codec show similar spectral and cepstral characteristics close to the original signal, in paticular for lower frequencies. The actual transmitted signal (TETRA Radio) attenuates low and high frequencies (compare Figure 4.16) with larger variations from the original signal for both spectral and cepstral values.

MIT Language Modelling Toolkit[9] to compute a trigram language model with modified Kneser-Ney smoothing. The language model is trained on about 150M running words of news wire and German newspaper texts. The vocabulary contains about 200k words. For decoding we use the Julius toolkit[10], which is typically used at Fraunhofer IAIS for such tasks due to its fast speech decoding for LVCSR.

Separation of Channel Influences

We conducted a set of experiments with the TETRA Corpus described in Section 3.2 with matching conditions for training and test set to separate the effects of the TETRA channel that lead to an ASR recognition performance drop. The evaluated conditions and the results in word error rates (WER) are given in Table 4.9.

The WER for the high quality broadcast data with 16kHz sample rate is at 26.6%. On the other end we achieve 42.3% WER for the TETRA Radio signal using actually transmitted speech data. In between these results we can attribute in absolute values an increase of 2.5% WER to the frequency low-pass effect using 8 instead of 16 kHz sampling rate. About another 6.5% WER can further be explained by the ACELP procedure within the TETRA encoding scheme. This can be witnessed when applying the conceptually similar AMR 4.75 Codec with the same bandwidth as TETRA to the 8 kHz speech data. Using the reference software implementation of the TETRA Codec instead of the AMR 4.75 only adds another 1.8% WER compared to the AMR 4.75. The particular processing inside the TETRA codec beyond the ACELP procedure does not seem to degrade the performance substantially. In the last step we can see a final 4.9% WER degradation between TETRA Codec and TETRA Radio, which can be attributed to the actual influence of the broadcast station.

In the last row of the table we can see the result for using acoustic models trained on the TETRA Codec training set but tested on the presumably similar TETRA Radio test data. We experience a surprising 20.5% rise in WER. Thus, there still must be some significant differences for both signals

[9] code.google.com/p/mitlm/
[10] julius.sourceforge.jp/

Condition train	Condition test	WER in %
Clean 16	Clean 16	26.6
Clean 8	Clean 8	29.1
AMR 4.75	AMR 4.75	35.6
TETRA codec	TETRA codec	37.4
TETRA radio	TETRA radio	42.3
TETRA codec	TETRA radio	62.8

Table 4.9: Performance loss through TETRA codecs. The step by step degradation of the ASR performance caused by different aspects of a TETRA coding and transmission is shown for matched conditions. Low-pass filtering, AMR coding, additional TETRA coding effects and actual hardware and transmission effects each increase the WER by about 2 to 7% absolute. The last row shows the effect of additional mismatch evaluating TETRA transmitted speech on TETRA coded speech without transmission.

probably caused by the TETRA hardware. Two differences that might explain these differences were discussed in Section 4.4.3. One major difference is probably caused by the harmonic distortion which cause certain critical distortion of the speech features, the other is a variation in cepstral mean normalisation caused by different lengths of the actual signal containing more non-speech frames at beginning and end in case of the TETRA Radio signal. Both effects probably explain a larger part of the 20.5% absolute drop in WER in mismatched conditions from TETRA Radio data tested on the TETRA Codec models compared to matched conditions with TETRA Radio models.We will further investigate these differences later in this work.

The main substitution errors for the TETRA radio test set evaluated on the TETRA codec models are based on phoneme confusion (substitutions) of /*i*/ and /*e*/ as well as /*m*/ and /*n*/ probably caused by the harmonic distortion of these rather harmonic sounds. Other errors include a deletion of phonemes like /*s*/ and /*d*/. However, the deletions of phonemes with significant parts of their energy in higher frequency bands are more likely erroneous because of the low-pass effect of TETRA radio already significantly attenuating frequencies above 2400 Hz.

Simulating hardware effects

In Table 4.9 we experienced already that acoustic models trained on data from the TETRA software codec do not perform very well on the more realistic TETRA Radio data. As discussed in the previous section this is probably caused by hardware and transmission effects including introduction of harmonic distortion and transmission channel noise caused by noise generated by the hardware as well as by errors during transmission.

We want to further investigate this gap of about 20.5% WER and also test the other acoustic models on our TETRA Radio test set. Table 4.10 shows the ASR performance for the different acoustic models tested with TETRA Radio data. The ideal case of no considerable mismatch is shown in the last row when training the acoustic models from data of exactly the same channel characteristics. This leads to a WER of 42.3%. We expect the performance of the other acoustic models to improve from top to bottom, as more and more aspects of TETRA coding and hardware effects are involved and thus the signal and the acoustic models can be expected to become more similar to the TETRA Radio data.

As opposed to our expectations we only achieve a minor performance boosts of 0.9% for AMR 4.75 and only 0.6% absolute for the TETRA Codec. Both software codecs hardly reduce the mismatch. The slightly lower WER in case of the AMR Codec compared to the TETRA Codec is probably an indicator that Motorola implemented the AMR 4.75 codec instead of the TETRA codec in the tested devices.

Training	WER
Clean 8	63.4
AMR 4.75	62.5
TETRA codec	62.8
TETRA radio	42.3

Table 4.10: ASR results for channel simulations tested on TETRA Radio data. WER increases significantly by more than 20% absolute in case of mismatched conditions when recognising TETRA transmitted speech (TETRA Radio) with acoustic models trained on simulations of the TETRA channel. AMR and TETRA Codec only slightly improve the results compared to the clean, low-pass filtered signal (Clean 8). These results provide the baseline in case that neither any TETRA hardware nor sufficient realistic TETRA Radio training data is available.

We now investigated the influence of artificially adding channel noise, equalisation effects and total harmonic distortion (see Table 4.11), this time by relying on the hardware equipment.

First, we checked the channel noise as introduced by the equipment, clearly audible as a buzzing sound when silence is transmitted. We recorded this "silence noise" and mixed it into the Clean 8 training material. This led only to a marginal reduction in WER of 0.3% absolute using the modified instead of the original Clean 8 acoustic models as we can see in Table 4.11. Next, we checked the influence of non-linear frequency responses in real transmission setups. Measuring the frequency response by recording a synthetic sweep from 0 to 4 kHz as shown in Figure 4.16, parameters for simulating the equalisation were obtained and applied to the training data. The WER compared to the unmodified data even increased slightly to 64.1%, so this effect apparently does not cause any large difference in WER either.

As discussed before it seems that the major contribution to the error rate is due to the harmonic distortion as witnessed in Figure 4.15. We simulated this effect by adding harmonic distortion to the clean signal as follows: for each bin b in the spectrogram, we have shifted its frequency content to the bins nb where $n \in \{2,3,\ldots,21\}$, i.e., we have added 20 harmonics to the original signal. The shifting was performed in the frequency domain (128 ms window length, 7/8-th overlap) and phase has been corrected such that phase angle ϕ in bin b has been transformed to phase angle $n\phi$ in bin nb. The strength of the individual harmonics has been calibrated by the intensity measured in a frequency sweep that has been transmitted via TETRA radio. This has led to much stronger harmonics than those found in speech signals transmitted via TETRA. Therefore, we have attenuated each harmonic by a factor of 0.02 which led to similar harmonic distortions as in real TETRA transmissions.

As can be seen in Table 4.11, the added harmonics improve the ASR results to 60.1% WER, an increase of 3.3% absolute. From all considered and simulated sources of distortion, this is the largest increase. This strongly suggests that harmonic distortion introduced by the equipment and probably further emphasised by the codec, contribute most to the large performance degradation. While for a human ear this effect might be inaudible, it heavily affects the MFCC balance. Manually checking the substitution errors indicate that most of them are indeed based on overtone confusion (e.g. the German "u" and the German "ü").

4.4.5 Conclusion

In general the performance of ASR significantly suffers from the channel characteristics of the TETRA channel compared to the high quality speech recordings.

We evaluated step by step certain aspects of the TETRA channel impact on the speech signal and the performance of ASR. We highlighted which aspect of the channel contributes most to the performance

Clean 8 with ...	WER in %
—	63.4
TETRA equipment channel noise	63.1
Equalization	64.1
Artificial harmonic distortion	60.1

Table 4.11: ASR results for TETRA Radio test data and simulated hardware effects. Additional hardware effects observed in Figures 4.15, 4.16 and 4.17 are simulated and evaluated. Equalisation and noise simulation do not yield any significant improvements. A simulation of the harmonic distortion reduces the WER by 3.3% absolute indicating a major impact on the ASR performance caused by mismatch due to additional harmonics.

drop. Interestingly, the hardware itself seems to have major influence on the WER due to harmonic distortion and further effects. It is interesting to note that, while the TETRA radio only seems to add a small error on a TETRA encoded signal when comparing the spectra, it is very hard to simulate the distortion of the station sufficiently. The supposedly little effect on the signal's spectrum shows major impact on the WER due to critical mismatch between the different realistic and simulated channels.

4.5 Conclusion

The influence of different sources of distortion on ASR are manifold and complex as we could demonstrate in this chapter. Generally, an integrated system designed for a certain task and well adapted in terms of the acoustic conditions and the domain yield the best possible results. We showed this in the introductory section developing such an integrated system for the task of command and control on the motorcycle. We could show that the training data and the training or adaptation of the acoustic models are crucial, but that other design decisions from hardware setup to the lexical knowledge further influence the results.

In a next step we concentrated on certain aspects of acoustic distortion that influence the quality of the recognition process. Every degradation of the signal from background noise to any distortion due to channel characteristics of microphones, hardware and transmission channels cause a drop in recognition performance. Often, these effects are non-linear and can even be phoneme dependent, for example in case of the Lombard effect or of different microphone technologies. Just for the major sources of distortion evaluated here the possible changes of the speech signal and the cepstral features can already be manifold and significant. A simulation or compensation of such non-linear, complex and sometimes even interconnected influences is very difficult and quite often not very successful. Background noise, for example, cannot just be considered to be additive as we could show. Generally, most approaches for simulation of background noise and distortion, including the presented additive noise and TETRA channel simulations, provide data still lacking certain aspects of the realistic signal. This is usually caused by wrong assumptions or simplifications and can have significant influences on the speech features and the ASR performance.

In all cases evaluated in this chapter, non-mismatched conditions between training and test data led to significantly better results than mismatched conditions even when applying common and successfully tested mismatch compensation techniques. While this is in line with other evaluations we could demonstrate and elaborate for the cases of background noise, differences in microphone channel characteristics, and the TETRA radio channel that even small differences in the signal and its spectrum can cause major difference in the cepstral coefficients used for recognition. Furthermore, we discussed and understood the dynamic characteristics and the variety of possible changes caused by these distortions.

Due to this complexity successful universal compensation techniques are difficult to develop. Common universal approaches are normalisation techniques aiming at normalising general statistics of the speech features. While these approaches usually improve the recognition results slightly, they will not be able to achieve a close to ideal compensation as more complex effects are not tackled at all.

We can further see from the results in this chapter that current speech features are far from being ideal for ASR as they incorporate a lot of information from the acoustic conditions as well as speech and speaker characteristics. Such information is completely irrelevant for a decision on the spoken content. Unfortunately, no considerably better speech features have been found yet making it necessary to compensate for this lack in the current features in one way or another. While feature normalisation and mismatch compensation approaches are one way as discussed before, another direction typically considered are well trained acoustic models. Whenever sufficient and preferably realistic training data can be gathered to train or adapt the acoustic models for all expected acoustic conditions, good results are also possible in more challenging and varying acoustic situations. A coverage of various conditions can be achieved, for example, by multi-conditional acoustic models with a sufficient number of Gaussian mixtures or by multiple acoustic models with algorithms successfully selecting the best set of acoustic models for the acoustic conditions of each utterance. Only in cases of unavoidable mismatch mismatch compensation should be applied as it usually does not yield comparably good results and in some cases might even introduce additional distortion.

Chapter 5

Blind Acoustic Model Selection

In the previous chapter we analysed and pointed out the different effects of additive noise, speech characteristics and channel effects on the acoustic features and the speech recognition process. Many approaches of robust ASR try to deal with one of these aspects only, often by making simplifications as, for example, the assumption that background noise is purely additive. While some of these approaches yield small to significant improvements whenever mismatch is present, they are still only coping with one of the many possible challenges. Other approaches like the ETSI robust front end use a cascade of compensation algorithms to deal with various distortions, mainly including additive noise and channel mismatch. These approaches are usually more universal and are often applied in cases of unavoidable mismatch. In general, the complexity of the possible influences on the features used for speech recognition limits the chances of successfully compensating all mismatch in realistic situations. We also discussed this in the previous chapter, where avoiding mismatch significantly outperformed compensation methods or a simulation of the distortion. While some time ago the lack of sufficient data was a much bigger problem, it became less and less critical in recent years. Even the rather expensive work in transcribing the data for machine learning algorithms usually pays of by providing significantly improved results compared to systems with mismatch between training and recognition.

Mainly two approaches exist to cover variations of speech features due to distortions in the acoustic models: training of one set of multi-conditional acoustic models with a sufficient number of Gaussian mixtures or training of multiple sets of acoustic models, each trained on a certain acoustic condition. While in noisy but rather homogeneous conditions in terms of noise characteristics multi-conditional acoustic models have shown comparatively good results, in situations where rather different acoustic conditions and distortions might be expected, multiple acoustic models might perform even better adapting more specifically to each of the various conditions. We will further investigate in this assumption and propose and evaluate a new multiple model approach in this chapter. The experiments in the previous chapter consolidated the statement that MFCCs are rather imperfect speech features for robust ASR, as information about the acoustic environment and channel conditions is inevitably present.We present a concept that tries to use this issue of MFCCs and other speech features to select the best acoustic models for each utterance based on this extra information. That means, we decide directly based on these speech features, which acoustic models from a set of models provide the lowest mismatch and should be the best acoustic models in terms of recognition accuracy for a given utterance.

Such an approach of a universal matching of speech features and acoustic models as proposed here is new, as to our knowledge only model selection approaches exist that concentrate on one of the sources of distortion only — typically focusing on speaker mismatch or environmental noise (for related work see Section 2.4.8). For the presented approach of blind acoustic model selection we do not consider at all, where the mismatch comes from. We use the mismatch information itself to classify, which of the available sets of acoustic models might work best.

In the following we will give a conceptual view on blind acoustic model selection as various approaches based on the general idea are possible. The concept is further motivated by discussing general

use cases and advantages of such a multi-model approach. We will have a close look at acoustic features and acoustic models discussing, why we believe that such an approach is possible and why the information in acoustic models can probably even be reduced for the model selection process. After presenting the general concept and its integration into an existing ASR system, we will present an approach for blind acoustic model selection including a way to speed up the selection process by using a compact model representation instead of the full sets of acoustic models. Furthermore, we present a way of compensating mismatch by relative cepstral normalisation based on our approach of blind acoustic model selection if mismatch is inevitable. In the final part of this chapter we will evaluate our approaches step-by-step and give a discussion on the results and potentials of the concept of blind acoustic model selection and relative cepstral normalisation.

5.1 The Concept of Acoustic Model Selection

Acoustic model selection is particularly interesting for multi-conditional ASR in scenarios with different but known acoustic conditions. Usually, multi-conditional acoustic models would be trained on representative training data, but considering the case of rather diverse acoustic conditions, ASR performance under a certain condition might decrease compared to single-conditional sets of acoustic models. Furthermore, such multi-conditional systems built on a single set of acoustic models need a full retraining of the acoustic models every time that new acoustic conditions must be covered by the ASR system.

As opposed to many other multi-model approaches (Section 2.4.8), our approach does not need any supervised preparation for a new set of acoustic models nor has any restrictions concerning the type of mismatch (e.g. additive noise only), as we directly use the acoustic information stored in the acoustic models. We do not need any prior knowledge about the acoustic conditions and we do not make any assumptions (except for assuming a constant type of acoustic features and acoustic models), so we call this approach blind acoustic model selection. While our approach is developed and evaluated based on hidden Markov models (HMMs) with mel-frequency cepstral coefficients (MFCCs) in this work, the general concept should be applicable to other types of acoustic models or features with no or only minor modifications.

One constraint, which is typical for any general multi-model approach, is the assumption that for each utterance a set of acoustic models exist that has no or only a low feature mismatch. This constraint also applies for our approach. If we take an utterance from unseen acoustic conditions, we might even be able to find the best matching acoustic models, but as the conditions are not covered by any of the sets of acoustic models, the recognition accuracy might still be very low. This problem also applies for multi-conditional acoustic models. Thus, an additional mismatch compensation step is still recommended for all cases, where we are not able to avoid any mismatch. In the context of this work a blind approach for compensation is preferred. The method should be based on general statistics or patterns without making further assumptions, for example, about the type of distortion causing the mismatch. Assuming that we find the best matching acoustic models for the presented features, mismatch compensation can probably benefit from using the information of the acoustic model selection step.

5.1.1 Advantages of Blind Acoustic Model Selection

Our approach to blind acoustic model selection provides a method that can easily be integrated in a standard ASR work flow and that allows an enhanced performance of the extended ASR system by "plug-and-play" of additional acoustic models. The algorithm is blind, and no particular acoustic characteristics (noise conditions, channel characteristics, etc.) are assumed for the acoustic models plugged

to the system to avoid limiting the approach to certain acoustic aspects as most other approaches for robust speech recognition do. Except for necessary restrictions — for example the type of speech features and the type of acoustic models to ensure comparability of features — no further requirements are defined.

An optimal ASR system incorporating blind acoustic model selection theoretically provides certain advantages:

- The system is able to cope with a broad range of acoustic conditions by multiple sets of acoustic models.
- The system can easily be extended to new acoustic domains by adding additional acoustic models for a new domain. (The performance on the previously supported acoustic domains is not reduced by this extension.)
- A set of acoustic models for each utterance is selected in terms of matching features, which is closely related to the acoustic recognition accuracy.
- The concept of acoustic model selection should enable the possibility of a fully automated, unsupervised way of adding acoustic models without requiring any additional manual or semi-automatic training of models like noise or concept models.

Nevertheless, blind acoustic model selection should not slow down the system dramatically during the recognition process. This means that the decoding process should be significantly faster than a multi-model multi-pass system performing full speech decoding for each of the acoustic models separately. Ideally, the overall speed should be close to a single model speech recognition system. Computationally expensive steps can still be used, but only in offline modules providing pre-calculated information for the online acoustic model selection. We will demonstrate in this work that we can benefit from some of these advantages in real-life applications, while other aspects need further investigations to generally improve ASR compared to other existing approaches.

5.1.2 Acoustic Information in Speech Features

Our approach builds on the assumption that speech features include sufficient information about the speech variabilities and acoustic distortion of their source signal.

We believe that **acoustic models contain sufficient information about speaker, channel and environmental characteristics** for a classification of the acoustic conditions.

Many approaches of ASR make use of acoustic models based on HMMs (Section 2.2.2). Usually a multivariate Gaussian function or multivariate Gaussian mixtures describe the probability distribution of possible observations of a state (Section 2.2.1), which can easily be defined by its means and variances. Means and variances of these probability distributions are calculated from training samples for each word or subword unit and are saved in the acoustic models.

During the recognition process the likelihoods of the HMMs are calculated from the observation using the probability distribution and transition probabilities. The sequence of words or subword units providing the maximum joint likelihood are presented as best hypothesis.

Unfortunately, the MFCCs of the same utterance recorded under different conditions or by a different speaker can be slightly or even quite different as we showed in Chapter 4. Thus, the means and variances of HMM based acoustic models trained on MFCCs (or other features) will be dependent on speaker and speaking characteristics, channel characteristics, environmental conditions etc. This leads to a mismatch

and a decreasing recognition performance of acoustic models from one domain on speech data of another domain.

On the other hand, the difference of means and variances thus inevitably provide information about the acoustic channel and environmental conditions the acoustic models were trained on. This is similar to MFCC-based speaker recognition and verification, for example, where also MFCCs are used as features to model each speakers characteristics (e.g., [132]).

We further believe that the **acoustic information in acoustic models can be reduced**, if only a classification on speaker, channel and environmental characteristics and no ASR should be done.

Considering that the purpose of a set of acoustic models in ASR is the recognition of words or subword units and not of the acoustic models themself, we should be able to reduce the amount of information necessary for model selection still keeping sufficient information about the major characteristics of the acoustic models. In their work on a multi-model approach in [95] the authors came to a similar conclusion also suggesting to consider a reduction of the acoustic information in future work.

If we consider only the mean values of a set of acoustic models, we can calculate the correlation matrix and the Euclidean distances between all mean vectors as shown for the MoveOn acoustic models trained on the noisy right microphone channel (see Section 3.3.4, Table 3.11, full training set) in Figure 5.1 and the Aurora 2 clean speech models in Figure 5.2. We can see that even beyond mixture components of a state or a HMM high correlation and low distances for many pairs of mean vectors indicate a high similarity between these vectors. This effect shows the similarities of several states of different but similar sounding phonemes. In case of the full word models of Aurora 2, we further have identical phonemes re-appearing as part of different word HMMs.

Instead of using all mean values of the acoustic models, which is usually not very handy for complex acoustic models like triphone models with tens of Gaussian mixtures per state, we will also consider to reduce the number of vectors for acoustic model selection to speed up the selection process. We assume that for discriminating two or more acoustic models, such a reduced set of features is usually sufficient (even though this might not be sufficient to discriminate between two subword units). Especially when considering a phrase or utterance level with tens or hundreds of feature vectors for decision, errors in a single frame introduced due to reduced information might hardly impact the averaged overall decision.

5.1.3 System Integration

The presented approach is based on the acoustic features extracted from an utterance, on the one hand, and on the information provided by the acoustic models, on the other hand, and can easily be integrated into a standard ASR work flow.

Figure 5.3 shows the integration of our approach (components with grey background) into an existing ASR system. A common system uses acoustic models and acoustic features extracted in a feature extraction step to calculate the best (acoustic) hypothesis of the spoken content in an acoustic decoding step of the ASR system. We suggest to add an acoustic model selection step that determines the best matching acoustic models for each utterance. Theoretically, this decision could be made using the entire acoustic information stored in a set of acoustic models, but especially for complex sets of acoustic models with triphone models and several mixtures per state (in case of HMMs) this can significantly slow down the recognition process. We suggest to use a compact representation of each set of acoustic models instead. The step of reducing the information in each set of acoustic models can be performed "offline" as soon as a new set of acoustic models is loaded. Thus, the selection process can usually be performed fast and online, while we enable more complex algorithms to extract a representative compact representation.

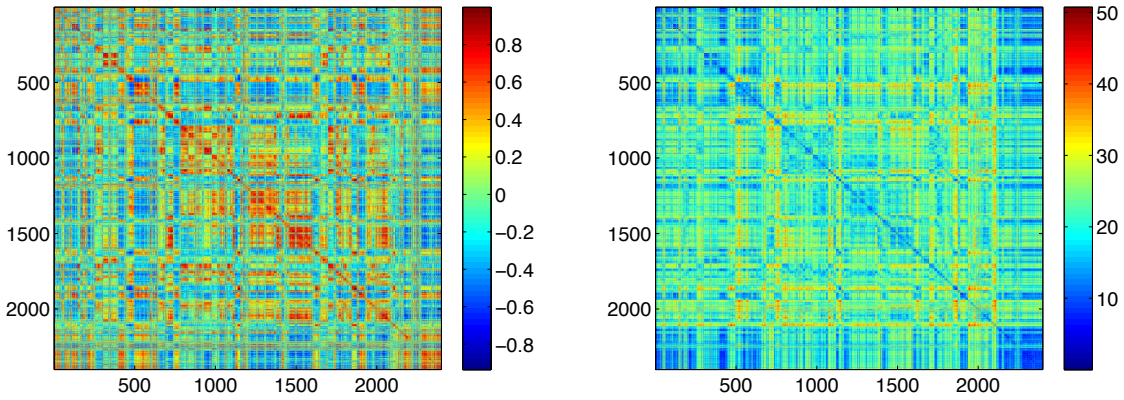


Figure 5.1: Correlation and Euclidean distance of mean vectors of acoustic models (MoveOn). The figure shows the correlation of (left) and Euclidean distance (right) between all mean vectors of the acoustic models of the MoveOn right close-talk channel. The values around the diagonal are correlations and distances between mixtures and states of the same HMM. High correlation and low distance are also visible in other areas indicating a high similarity of many mean vectors even beyond the mixtures of an HMM.

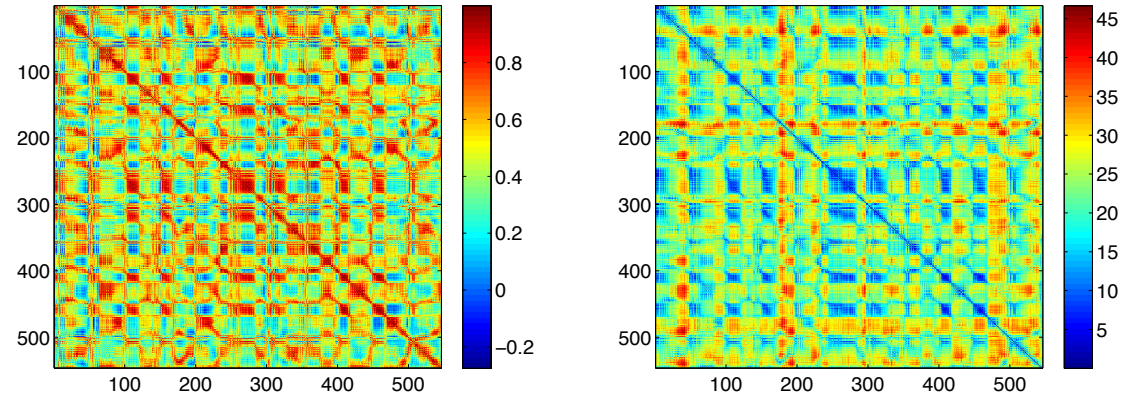


Figure 5.2: Correlation and Euclidean distance of mean vectors of acoustic models (Aurora 2). The correlation of (left) and Euclidean distance (right) between all mean vectors of the clean speech Aurora 2 acoustic models (similar to 5.1) also indicate a high similarity of many mean vectors beyond the mixtures of an HMM. Here in particular the full word models introduce additional similarities as certain phonemes are present in several of the modelled words.

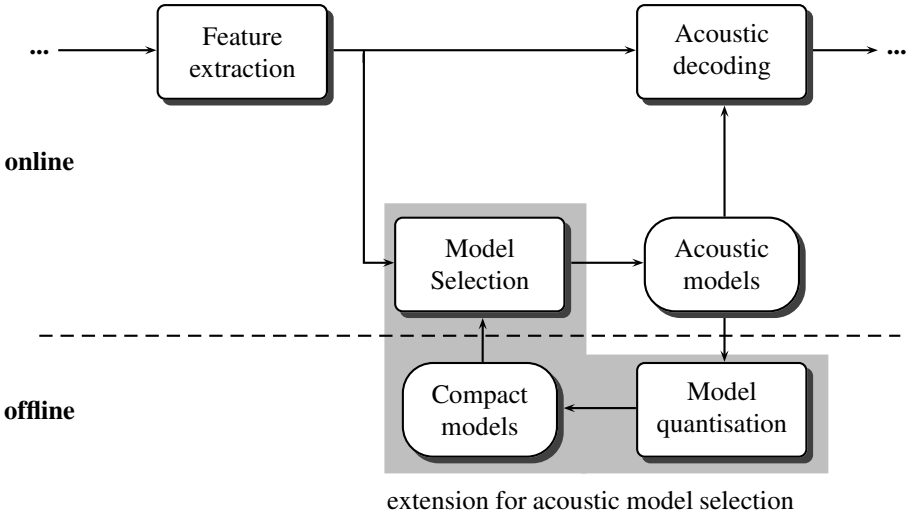


Figure 5.3: Integration of blind acoustic model selection into ASR work flow. The concept of acoustic model selection (grey components) can easily be added to a standard ASR system. An offline model quantisation reduces the dimension of each set of acoustic models providing compact models. During the actual recognition process the extracted speech features and the compact models are used for acoustic model selection. The full set of acoustic models related to the best matching compact model is selected for acoustic decoding during ASR.

5.2 Codebook-based Acoustic Model Selection

The concept of blind acoustic model selection should be applicable to different types of features and approaches used for ASR, but we will focus on a common state of the art system with MFCCs and HMMs with multivariate Gaussian distributions in this work.

Our approach is based on the mean values of the acoustic models only, which already provide a high degree of information. Transitions, weights, variances, and other information are neglected. In the conclusion of [95] the authors suggested that the detailed phonetic information they used for speaker clustering might be ignored and that vector quantisation based information could be sufficient to characterise a speaker. As motivated by our observations in Section 5.1.2 we further generalise this assumption, so that vector quantisation based information is assumed to be sufficient to characterise any acoustic conditions beyond speaker characteristics within our speech features.

Thus, we use vector quantisation with K-means to extract a compact representation of the acoustic models before we apply a distance based classification. In the following we will detail this approach of model quantisation and selection.

5.2.1 Compact Representation of Acoustic Models

Even though a matching with full sets of acoustic models is possible, it is not very efficient in terms of computation time, especially for complex acoustic models. Thus, we aim at extracting a compact representation of a set of acoustic models, which should reduce the computational load but provide comparable results to an approach using the full set.

In a first step we extract only the mean vectors of a set of acoustic models without considering the context (HMM, state, mixture). To reduce the complexity of the distance calculation we further reduce the number of mean vectors extracted from the acoustic models. We will show later that this practically

does not influence the model selection process compared to using a full set of means. One way of reducing the number of means is K-means clustering. We assume that a quite high compression of acoustic models is possible as high correlations and low distances between certain mean vectors within a set of acoustic models are common as we could see in Section 5.1.2. Such small differences in the feature vectors even beyond HMMs are probably not crucial for acoustic model selection, even though they are relevant for discriminating similar sounding phonemes during ASR.

A a family of acoustic models M can then be described as a set of mean vectors $M = \{m_1, m_2, \ldots, m_S\}$. We now want to find a compact representation $\widetilde{M} = \{\mu_1, \mu_2, \ldots, \mu_K\}$ of M with $K < S$. We use K-means with Euclidean distance to cluster M into K clusters with the centroids μ_k, which define our compact representation $\widetilde{M}$. Practically, our compact representation $\widetilde{M}$ could also be considered as codebook of a vector quantisation of M. As K-means minimises the sum of (squared) distances to the cluster centres, the introduced mean squared error is minimal when quantising M with codebook $\widetilde{M}$, i.e. when using the compact representation instead of the full acoustic models in a distance based classification as elaborated in the following subsection.

5.2.2 Acoustic Model Selection

For acoustic model selection various classification approaches are possible. In our approach we use minimum distance classification to decide on the best set of acoustic models used for ASR. Therefore, we must determine the distance between each set of acoustic models and the utterance. As an utterance contains several frames, we first estimate the distance of each frame's feature vector x_n to each vector m_s of the set of acoustic models or each vector μ_k of its compact representation. In a second step we calculate the overall distance of an utterance to each set of models (or its compact representation) by averaging the minimum distances of each frame's vector to the set of models (or its compact representation). In the following we will only use the terminology for the compact representation, while the same calculation for the full set of mean vectors is equivalent.

In detail, we calculate the distance between an utterance and the compact representation of each set of acoustic models on a frame by frame basis. The minimum distance of a frame to a set of acoustic models is defined by the distance of the feature vector of a frame to the nearest neighbour in the compact representation. We will compare two distance measures for estimating the distance $d(k,n)$ between a frame n with its feature vector x_n and the k-th centroid μ_k of a compact representation $\widetilde{M}$. The first is Euclidean distance (Equation 5.1), which is well known and determined straightforward:

$$d_2(k,n) = \sqrt{\sum_i \left(\mu_{k,i} - x_{n,i}\right)^2} \tag{5.1}$$

The second is Mahalanobis distance (Equation 5.2), which also considers differences in dynamic range of the feature vectors' dimensions:

$$d_M(k,n) = \sqrt{\left(\mu_k - x_n\right)^T \Sigma^{-1} \left(\mu_k - x_n\right)} \tag{5.2}$$

We use a global covariance matrix Σ with the global average of each matrix element. As the acoustic models considered in our approach have diagonal covariance matrices, the global matrix is also diagonal.

Furthermore, we define the overall distance[1] $d(U||\widetilde{M})$ between an utterance $U = \{x_1, x_2, \ldots, x_N\}$ and the compact representation of a set of acoustic models $\widetilde{M}$ by the mean of the minimum distances of all N frames of the utterance:

[1] Strictly speaking, this is not a true distance metric any more, as it is not necessarily symmetric.

$$d(U||\widetilde{M}) = \frac{1}{N}\sum_{n=1}^{N}\min_{k}(d(k,n)) \tag{5.3}$$

We select the acoustic models M with the compact representation $\widetilde{M}$ to which the utterance U has minimum distance. In case of Euclidean distance this is equivalent to selecting the codebook $\widetilde{M}$, which minimises the mean squared quantisation error $d(U||\widetilde{M})$ for utterance U.

For comparison we also include a maximum-likelihood selection based on Equation 2.9 for acoustic model selection. Here, we consider all multivariate Gaussian distributions with means and variances of the full set of acoustic models AM_i. We ignore the concept of Gaussian mixtures and treat all S pairs of mean vectors and variances (m_s and Σ_s) of a set of models as single multivariate Gaussian distributions. For each feature vector x_n of an utterance U and Gaussian distribution we can calculate the probability:

$$b_s(x_n) = \frac{1}{(2\pi)^{K/2}|\Sigma_s|^{1/2}} e^{-\frac{1}{2}\left((x_n-m_s)^T\Sigma_s^{-1}(x_n-m_s)\right)} \tag{5.4}$$

For each set of acoustic models AM_i we determine the probability $P(U|AM_i)$ for U given the set of all Gaussian distributions of AM_i. Therefore, we calculate and join the probabilities for all frames x_n and all S distribution functions of AM_i:

$$P(U|AM_i) = \prod_{n=1}^{N}\frac{1}{S}\sum_{s=1}^{S}(b_s(x_n)) \tag{5.5}$$

We do not include any state or transition information as we do not consider any particular phoneme or word models, but only means and variances of all distributions in a set of acoustic models. Even though this is not identical to the acoustic decoding of the speech recogniser, selecting the set of acoustic models AM_i that maximises this likelihood function $P(U|AM_i)$ is closely related to the probabilistic acoustic decoding. As this calculation is not very efficient and requires a full set of means and variances from the acoustic models, we introduce it for comparison only. The Mahalanobis distance closely related to the likelihood is not convenient here, as it only considers the exponent but neglects the factor, thus, favouring Gaussian distributions with small variances (in case of a diagonal covariance matrix as used here) compared to the full probability value. Thus, we only use the Mahalanobis distance to compensate for different variances of the feature vectors dimensions as introduced in Equation 5.2.

5.3 Recombination of Acoustic Models

Alternative to a selection from several sets of acoustic models we also evaluate a recombination of the separately trained sets of acoustic models. Thus, the final selection of the appropriate HMM is done by the speech recogniser during the acoustic decoding process with maximum-likelihood Viterbi decoding. We consider two different approaches for a recombination, the combination of HMMs and the concatenation of HMMs.

5.3.1 Combination of HMMs

In a first setup we combine the acoustic models on an HMM level. Each of our domain-specific acoustic models uses the same phonemes and concept of MFCCs (with 39-dimensional feature vectors calculated as detailed in Section 3.3.6) and Gaussian mixtures describing each state. Similar to multi-conditional acoustic models we will extend each HMM to a higher number of Gaussian mixtures by concatenating

the mixture components of all sets of acoustic models and adapting weights and transition probabilities for each mixture and HMM appropriately. We weight all sets equally, so that the resulting weights $\tilde{w}_{s,\ell} = 1/L w_{s,\ell}$ of each mixture are simply calculated by dividing the existing weights $w_{s,\ell}$ by the number L of sets that are joined. The new transition probabilities $\tilde{a}_{ij}$ for each HMM are calculated as the mean of the L transition probabilities $a_{ij,\ell}$ of the joined models:

$$\tilde{a}_{ij} = 1/L \sum_{\ell=1}^{L} a_{ij,\ell} \tag{5.6}$$

The resulting structure of the set of acoustic models is similar to the multi-conditional acoustic models but with each mean vector and covariance matrix more specifically trained on one of the acoustic conditions.

5.3.2 Concatenation of HMMs

In the previous approach we lose the specific transition probabilities of the acoustic models, as we finally combine the L different transition probability matrices in the separate acoustic models — each for a specific acoustic condition — into a single matrix of the combined HMM. In particular for speech variabilities caused by different speakers, accents, or the Lombard effect, the transition probabilities can be rather different for acoustic models specifically trained for one or another acoustic condition or speaker. Thus, we also test another approach of recombining the specific sets of acoustic models into one set of models. To preserve the specific transition probabilities we do not change the HMMs in the set of acoustic models, but increase the number of HMMs in the combined set of models by the factor L (as we have the same number of phonemes in each of our L sets of specific acoustic models). That means that we just concatenate all HMMs of the different acoustic models in one general set of acoustic models. In case of phoneme recognition each phoneme is now represented by L different HMMs with each of these HMMs specialised on a certain acoustic condition. Thus, again we let the speech recogniser decide about the maximum-likelihood path. This also enables paths connecting HMMs of one acoustic condition to HMMs trained on another acoustic condition making it more flexible in case of changing acoustics within one utterance or segment compared to our proposed blind acoustic model selection approach. Finally, the resulting HMM sequence is mapped back to the standard phoneme alphabet.

5.4 Relative Approach for Mismatch Compensation

While avoiding mismatch by using matching acoustic models generally yields the best results in ASR, mismatch cannot be avoided in all cases, so that an ASR system will face some unknown acoustic conditions at some point. Thus, we adapt two existing, successfully tested approaches for cepstral normalisation to perform a relative normalisation incorporating information from our concept of acoustic model selection.We do not want to limit our approaches to certain conditions, so we neither assume any particular acoustic condition or type of mismatch[2] nor any knowledge about the spoken utterance or subword unit for our compensation approaches. These are strong limitations as we demonstrated in the previous chapter about acoustic distortion, as acoustic mismatch can depend significantly on the type of phoneme as well as the type of distortion. In the following we present two approaches of relative

[2] Even though we do not assume any particular mismatch, the presented approaches are conceptually limited to compensate for certain linear effects on cepstral features only. If such linear mismatch is not present, they ideally do not change the features at all.

cepstral normalisation that aim at a normalisation of the cepstral values of the feature vectors of an utterance relative to the selected acoustic models. Both approaches focus on a normalisation of mean and gain-related cepstral characteristics.

The presented approaches can also be applied without blind acoustic model selection as general normalisation steps. In this case a codebook for the single set of acoustic models is created before the recognition as described in Section 5.2.1 and provided for normalisation. During the recognition we then determine for each frame of an utterance the nearest neighbour in this codebook and apply the relative normalisation as presented here.

5.4.1 Relative Cepstral Normalisation

In Section 2.4.3 we already discussed the approach of quantile-based cepstral dynamics normalization (QCN - [70]). This approach is related to CMN and CGN described in Section 2.4.1, which are often applied in ASR systems as they successfully compensate some of the mismatch caused by additive noise and channel distortions. Still, such approaches have two disadvantages. First, acoustic models must already be trained on normalised features using the same approach. Second, during online recognition an utterance can only be normalised on an utterance level, which can lead to a normalisation on a different basis as the phoneme distribution within the utterance can be quite different to the one in the training data of the acoustic models. As these approaches are still successful, we will present two new approaches trying to compensate for these disadvantages by using a relative normalisation.

In our approach for acoustic model selection we determined the nearest neighbour from all mean vectors of a set of acoustic models for each frame of an utterance (Section 5.2.2). Instead of a mean and gain normalisation on the training data and the test utterance, we are also able to perform a normalisation of the mean and gain of the utterance relative to the mean and gain of the nearest neighbours of the acoustic models now. The goal is to reduce mismatch due to different normalisation bases and to enable normalisation also in case of unnormalised acoustic models.

The first approach builds on the statistical mean and gain differences, while the second one uses the averaged individual cepstral mean and gain differences.

Normalisation on General Statistics

In this first approach we consider the general statistics of the feature vector set of the frames of the utterance and the nearest neighbours to these vectors in the set of acoustic models (or its compact representation). This results in T feature vectors of the test utterance $X = \left(c_1^{(X)}, c_2^{(X)}, \ldots, c_T^{(X)}\right)$ and the same number of vectors from the set of acoustic models $Y = \left(c_1^{(Y)}, c_2^{(Y)}, \ldots, c_T^{(Y)}\right)$. For both sequences of vectors X and Y we can calculate the mean $\overline{c}_n^{(X)}$ and $\overline{c}_n^{(Y)}$ as well as the gain $g_n^{(X)}$ and $g_n^{(Y)}$ for each cepstral coefficient n:

$$\overline{c}_n = \frac{1}{T}\sum_{t=1}^{T} c_{n,t}\,, \tag{5.7}$$

and

$$g_n = \max_t\left(c_{n,t}\right) - \min_t\left(c_{n,t}\right). \tag{5.8}$$

We could now normalise mean and gain based on the calculated value, using the resulting cepstral values $\tilde{c}_n, i$ of the following relative normalisation for recognition:

$$\tilde{c}_{n,i}^{(X)} = \frac{g_n^{(Y)}}{g_n^{(X)}} \left(c_{n,i}^{(X)} + \left(\overline{c}_n^{(Y)} - \overline{c}_n^{(X)} \right) \right). \tag{5.9}$$

Unfortunately, we cannot expect that the nearest neighbour $c_n^{(Y)}$ of $c_n^{(X)}$ necessarily represents the correct state of the correct HMM even though they have minimum distance. Thus, we introduce three steps to reduce the effects of outliers and to avoid "unfavourable pairs" (minimum distance pairs that do not represent the same state of the HMM) for relative normalisation:

1. We only consider "reliable pairs" of vectors $Y_r = \left(c_1^{(Y_r)}, c_2^{(Y_r)}, \ldots, c_M^{(Y_r)} \right)$ to $X_r = \left(c_1^{(X_r)}, c_1^{(X_r)}, \ldots, c_M^{(X_r)} \right)$ as subsequence of Y and X with $M \leq N$ and reordering of indices. We define reliable pairs as all pairs of vectors with a distance below a threshold Θ, as we assume that large distances indicate unfavourable pairs.

2. We further introduce two weights w_1 and w_2 to influence the level of normalisation of cepstral mean and gain. Both weights can have a value between 0 and 1, with 0 resulting in an unmodified cepstral vector and 1 in a full normalisation. Weights close to zero reduce the level of introduced distortion in case of many unfavourable vector pairs, but also reduce the normalisation effect in case of mainly favourable pairs.

3. Instead of cepstral mean and gain normalisation we use quantile-based cepstral dynamics normalisation (QCN - [70]). As mentioned by the authors, this approach is less sensitive to outliers than the overall minimum and maximum defining the gain.

Similar to the mean and gain normalisation as described above, we adapt QCN to a relative normalisation. We further introduce the two weights and consider reliable pairs only. QCN is usually performed as follows based on the quantiles $q_j^{(c_n)}$ and $q_{100-j}^{(c_n)}$ (compare Section 2.4.3, Equation 2.43)

$$c_{n,i}^{QCN_j} = \frac{c_{n,i} - \left(q_j^{(c_n)} + q_{100-j}^{(c_n)} \right)/2}{q_{100-j}^{(c_n)} - q_j^{(c_n)}} \tag{5.10}$$

In this equation we have the centring term $(q_j^{(c_n)} + q_{100-j}^{(c_n)})/2$ and the normalisation factor $1/(q_{100-j}^{(c_n)} - q_j^{(c_n)})$. To estimate the relative normalisation, we calculate the quantiles $q_j^{(c_n)}$ and $q_{100-j}^{(c_n)}$ for each cepstral value and each of the sets of the reliable pairs X_r and Y_r independently. Based on these quantiles we determine a weighted compensation term for relative centring b_n and a weighted factor for relative dynamics normalisation f_n for each cepstral coefficient:

$$b_n = \frac{w_1}{2} \left(\left(q_j^{(c_n^{(Y_r)})} + q_{100-j}^{(c_n^{(Y_r)})} \right) - \left(q_j^{(c_n^{(X_r)})} + q_{100-j}^{(c_n^{(X_r)})} \right) \right), \tag{5.11}$$

and

$$f_n = \frac{w_2 \left(q_{100-j}^{(c_n^{(Y_r)})} - q_j^{(c_n^{(Y_r)})} \right) + (1 - w_2) \left(q_{100-j}^{(c_n^{(X_r)})} - q_j^{(c_n^{(X_r)})} \right)}{q_{100-j}^{(c_n^{(X_r)})} - q_j^{(c_n^{(X_r)})}}. \tag{5.12}$$

With f_n and b_n we can normalise the cepstral features of the utterance relative to the reference vectors of the set of acoustic models:

$$\tilde{c}_{n,t}^{(X)} = f_n \left(c_{n,t}^{(X)} + b_n \right). \tag{5.13}$$

The normalised values $\tilde{c}_{n,t}^{(X)}$ are used instead of the original cepstral vectors for recognition with the selected acoustic models. As our approach is a relative approach of QCN, we call it *relative QCN* (rQCN).

Normalisation on Averaged Individual Characteristics

In the previous section we used general statistics of reliable pairs for normalisation. Another approach to derive a mean and gain normalisation is to consider each individual pair of vectors. Again, we use reliable pairs and weights to reduce the effect of outliers as introduced in the previous section. Instead of the general statistics of the distribution not considering any particular relation between each of the pairs, we now directly use the differences of the pairs to estimate the statistics of these differences to normalise accordingly. With the assumption that both vectors can be considered to be observations generated by the same state of an HMM (which ideally should be the case) we can directly determine the differences in mean and gain to normalise the features. As this assumption cannot be assumed to be valid in all cases, even when considering so called reliable pairs (see Section 5.4.1) only, this approach is presumably more prone to outliers caused by unreliable pairs than rQCN.

First, we determine a weighted additive term b_n reducing the mean offset between each vector pair. This term is calculated as the average difference of a cepstral coefficient over all reliable pairs. This is equivalent to a (weighted) relative cepstral mean normalisation, as the following equation based on the notation from the previous section shows:

$$b_n = w_1 \left(\frac{1}{M} \sum_{m=1}^{M} \left(c_{n,m}^{(Y_r)} - c_{n,m}^{(X_r)} \right) \right) = w_1 \left(\frac{1}{M} \sum_{m=1}^{M} c_{n,m}^{(Y_r)} - \frac{1}{M} \sum_{m=1}^{M} c_{n,m}^{(X_r)} \right) = w_1 \left(\overline{c}_{n,m}^{(Y_r)} - \overline{c}_{n,m}^{(X_r)} \right), \tag{5.14}$$

In a second step similar to cepstral gain normalisation we estimate an additional compensation factor. This compensation factor f_n is calculated from the vectors of the reliable pairs weighted by w_2 similar to the previous approach:

$$f_n = \underset{m}{\text{median}} \left(\frac{w_2 c_{n,m}^{(Y_r)} + (1 - w_2) \left(c_{n,m}^{(X_r)} + \overline{\Delta c}_n \right)}{c_{n,m}^{(X_r)} + \overline{\Delta c}_n} \right) \tag{5.15}$$

We use the median instead of the mean value, as large feature differences in unfavourable pairs can create large factors shifting the mean significantly, but do not change the median (as long as only a few outliers occur). Then the features of an utterance are normalised by removing the weighted mean offset b_n and adjusting the gain with factor f_n equivalent to Equation 5.13 resulting in the normalised features $\tilde{c}_{n,t}^{(X)}$ for recognition. In this work we refer to this approach as *relative CGN* (rCGN).

5.5 Evaluations

In the following we evaluate our approaches for blind acoustic model selection and relative mismatch compensation. We analyse the effect of the compact representation on the classification and speech recognition performance and discuss the advantages and disadvantages of our approaches.

5.5.1 Evaluation Setup

In our evaluations we use the three speech corpora introduced in detail in Chapter 3, namely the Aurora 2 evaluation corpus as commonly used reference corpus, the MoveOn Corpus for evaluation on realistic noisy conditions, and the TETRA Broadcast Corpus representing a LVCSR task with complex triphone models. We use HTK for training and recognition except for the LVCSR task where recognition is performed using the Julius speech recognition decoding software, which is typically used for LVCSR at Fraunhofer IAIS due to its fast decoding speed for this task.

Set I: MoveOn Corpus

We use the full training and test sets of the throat microphone and the right microphone channel of the MoveOn Corpus (Section 3.3). For each channel set we consider the office subset (also referred to as "clean speech") and the motorcycle subset (also "noisy speech") separately resulting in four evaluation subsets. Each subset contains a disjoint training and test set. For each of the four subsets acoustic models are trained on the respective training set.We further train two multi-conditional acoustic models on all available training data of both channels.

We use HMM-based acoustic models trained on MFCCs (with 39-dimensional feature vectors as described in Section 3.3.6). The acoustic models are monophone models for phoneme based recognition with 44 phonemes plus silence and short pause. Each acoustic model consists of three states modelled by 16 Gaussian mixtures. We train two multi-conditional acoustic models, one with 16 Gaussian mixtures and one where we increased the number of mixtures to 64 to match the sum of mixtures of all sets of domain-specific models.

Set II: Aurora 2

We only consider test set A of the Aurora 2 evaluation set (Section 3.1), which contains subsets with noise from the domains subway, babble, car, and exhibition added with different signal-to-noise ratios (SNRs). We use the clean and the multi-conditional training sets defined in the corpus. Furthermore, we divide the multi-conditional training set (8440 utterances) into further training subsets with each subset (1688 utterances) containing speech data from only one of the four noise domains or clean speech. All test sets contain 1001 utterances.

We use the standard Aurora 2 evaluation setup for HMM training and standard feature extraction with HTK, except for the MFCCs where we also apply cepstral mean normalisation. One HMM for each of the digits from "one" to "nine", "zero" and "oh" as well as an additional model for silence and for short pause is trained. We use 16 (modelled) states with three Gaussian mixtures per state, except for a second set of multi-conditional acoustic models with 15 mixtures per state (matching the total number of mixtures of all five domain-specific sets).

Set III: TETRA Corpus

Our third evaluation set represents the complex task of LVCSR and is assembled from our TETRA Corpus described in Section 3.2. We use the 8 kHz clean speech training and test data ("Clean 8") and the speech data transmitted via TETRA ("TETRA radio") for the acoustic model selection evaluation. For evaluation of our mismatch compensation approaches we further include acoustic models from 8 kHz clean speech, AMR 4.75 codec and TETRA codec, which are tested with the mismatched TETRA radio test set.

For each of the sets we train a separate set of acoustic models from the respective training data. We use HTK for feature extraction of a set of g features (12 MFCCs with short time energy, plus deltas and accelerations). Cepstral mean normalisation is used. We train triphone acoustic models with 32 Gaussian mixtures per state resulting in a set of acoustic models with about 200,000 mean and variance vectors. The same trigram language model as briefly described in Section 4.4.4 is used for recognition.

The TETRA Corpus is prepared for complementary evaluations showing the capabilities and effects in case of a more complex LVCSR task. Thus, we only use this corpus for some final evaluations on blind acoustic model selection and relative cepstral normalisation but not for preliminary evaluations as opposed to the MoveOn and the Aurora Evaluation Sets.

5.5.2 Codebook-based Acoustic Model Selection

Our approach of blind acoustic model selection incorporates a reduction of the information of each set of acoustic models and the actual selection approach. We first evaluate the influence of the information reduction on the frame classification error and speech recognition accuracy in acoustic model selection when using a codebook instead of the full set of acoustic models. Afterwards, we will analyse the selection performance and confusion using our approach of blind acoustic model selection and compare it with multi-conditional training of acoustic models.

Effect of Compact Representation

Considering the effect of the compact representation of the acoustic models, the first measure of interest is the vector classification error when classifying each mean vector $m_{i,s}$ of the set of mean vectors M_i of the i^{th} set of acoustic models by minimum Euclidean distance to the cluster centers defined by the compact representation $\tilde{M}_j$ of the j^{th} set of acoustic models. Each assignment of a mean vector $m_{i,s}$ to a compact representation $\tilde{M}_j$ with $j \neq i$ is counted as classification error. A second measure particularly interesting for our task is the effect on the ASR performance with acoustic model selection using compact representations of different dimensions. Therefore, we compare the ASR results of our acoustic model selection approach for a full set of mean vectors M and several compact representations $\tilde{M}_k$ of different dimensions K.

Figure 5.4 shows the maximum classification errors for the MoveOn and the Aurora acoustic models. Furthermore, the performance of the ASR system using blind acoustic model selection with the full set of acoustic models as well as with compact representations of different dimensions is presented in the same figure. The full set of acoustic models has 2400 (MoveOn) or 546 (Aurora) mean vectors. We see that the classification error increases significantly for low values of K — up to an error of 40% (MoveOn) and 69% (Aurora) for $K = 5$. Still, the classification error hardly affects the acoustic model selection, as the ASR performance with acoustic model selection stays almost constant down to a K of about 50 to 100. As we calculated the classification error for a frame classification and we use a matching of full utterances (with usually hundreds of frames) in our implementation, scattered classification errors have no major impact during acoustic model selection. This explains the constant ASR performance compared to the increasing vector classification error. Thus, we can reduce the acoustic models to a significantly smaller codebook for acoustic model selection without losing accuracy during ASR. This enables a much faster computation for online acoustic model selection. For further evaluations we use $K = 300$, which shows a vector classification error just below 10% for both evaluation sets.

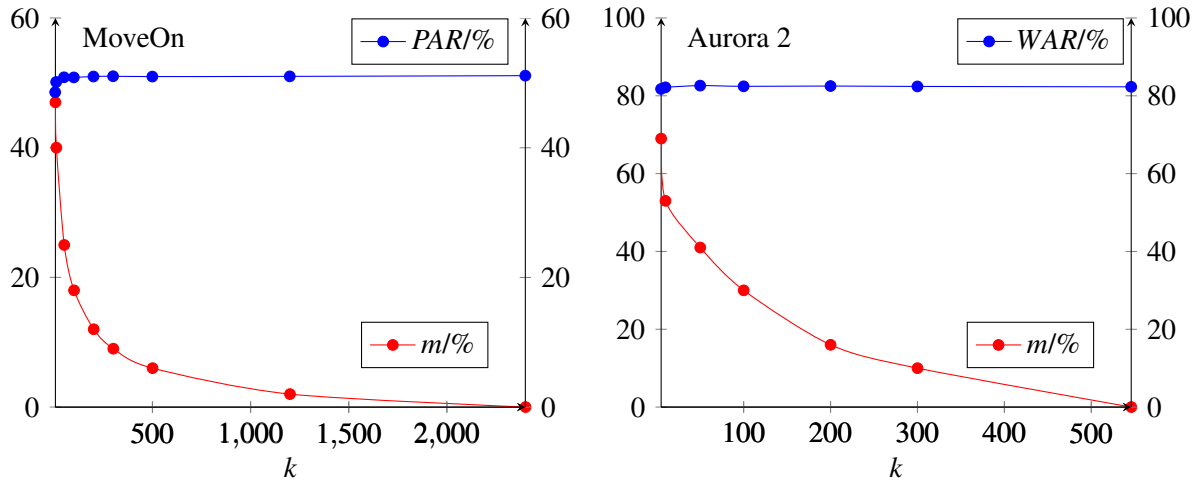


Figure 5.4: Vector misclassification and recognition accuracy for blind acoustic model selection. The figure shows the percentage m of incorrectly assigned mean vectors when classifying each vector of the full set of models using the compact representations with k clusters for MoveOn (left) and Aurora 2 (right). Furthermore, phoneme accuracy rates (*PAR*) or word accuracy rates (*WAR*) are provided. While the misclassification of single frames increases with decreasing k, the recognition accuracies are hardly affected.

Selection Performance and Confusion

We determine for which percentage of utterances each set of acoustic models performs best in terms of phoneme accuracy (MoveOn) or word accuracy (Aurora). We can see the statistics for the four sets of acoustic models of the MoveOn Corpus in Table 5.1, left-hand side. The values along the diagonal show the expected tendency that a set of acoustic models trained on the same matched conditions usually performs best for most utterances. But we can also see that there is a considerable amount of utterances in each set, where another set of models performs equally well or even better. Especially for the throat microphone signal we can see that the data recorded in noisy environments is often well recognised with the acoustic models trained on clean speech data from the office environment and vice versa. As mentioned in Section 3.3.3, the throat microphone does not pick up airborne sound and is expected to be less affected by environmental noise, which might explain this effect as the microphone channel is identical and only the environmental noise and related effects might differ.

But also for the close-talk microphone channel some similarity for both acoustic conditions can be seen. This can be explained by three influences: First, the acoustic channel is identical, so that this source of mismatch is avoided. Second, the motorcycle recordings contain training and test data, which in some cases can be considered as almost clean speech. Thus, the mismatch between motorcycle recordings and office recordings hardly suffers from additive noise in such cases. Third, the motorcycle recordings offer much more training data, which enables improved modelling of speech, which also improves the results for the office test data.Interestingly, noisy throat microphone data tested on noisy close-talk models and vice versa perform rather well for about 10% of the test utterances. This is probably an indication of the Lombard effect, which is one of the few sources of distortion, which both training and test sets have in common, and thus, are modelled in both sets of acoustic models.

Table 5.1, right-hand side, shows the relative confusion matrix for blind acoustic model selection for the MoveOn evaluation sets. We can see that the acoustic models from the same domain are usually preferred by our selection approach. Confusion is higher for acoustic models which rather often provide equally good or better results than the domain-specific models. The effect of using a compact repres-

	Acoustic models					Acoustic models			
is best	tc	tn	rc	rn	**k=300**	tc	tn	rc	rn
tc	70%	37%	17%	4%	tc	1.00	0.00	0.00	0.00
tn	28%	83%	5%	6%	tn	0.09	0.91	0.00	0.01
rc	5%	4%	85%	21%	rc	0.01	0.01	0.96	0.03
rn	3%	11%	8%	93%	rn	0.00	0.01	0.01	0.98

Table 5.1: Confusion matrix for blind acoustic model selection (MoveOn). The left-hand side of the table shows the percentage of utterances of each test set (tc: throat clean, tn: throat noisy, rc: right clean, rn: right noisy) for which the set of acoustic models provide the best phoneme accuracy rates. Acoustic models without mismatch (diagonal) usually perform best. The right-hand side shows the confusion matrix for blind acoustic model selection and a compact representation of size $k = 300$. Confusion mainly occurs in case of acoustic models providing rather high values in the left part of the table.

	Acoustic models (is best)						Acoustic models (confusion)				
N1, best	clean	N1	N2	N3	N4	**N1, 300**	clean	N1	N2	N3	N4
clean	99%	64%	81%	77%	49%	clean	1.00	0.00	0.00	0.00	0.00
20 dB	86%	97%	95%	94%	96%	20 dB	0.01	0.84	0.03	0.01	0.11
15 dB	64%	96%	91%	89%	94%	15 dB	0.00	0.88	0.02	0.00	0.10
10 dB	27%	94%	84%	83%	91%	10 dB	0.00	0.89	0.03	0.02	0.06
5 dB	9%	89%	70%	68%	84%	5 dB	0.00	0.77	0.07	0.04	0.12
0 dB	6%	80%	50%	56%	70%	0 dB	0.00	0.62	0.22	0.12	0.03
-5 dB	19%	66%	43%	56%	62%	-5 dB	0.00	0.21	0.16	0.58	0.05

Table 5.2: Confusion matrix for blind acoustic model selection (Aurora 2). The left-hand side of the table shows the percentage of utterances of each test set (clean and N1 to N4 of different SNRs) for which the set of acoustic models trained on noisy data from the subway domain (N1) provides the best word accuracy rates. Acoustic models without mismatch (clean–clean and N1–N1) usually perform best. In case of noise, the acoustic models trained on the noise domain N1 also performs quite well on test data with noise types N2 to N4. The right-hand side shows the confusion for blind acoustic model selection and a compact representation of size $k = 300$. Confusion mainly occurs for low SNRs.

entation of the acoustic models slightly increases the confusion, especially for rather low dimensions of 50 and smaller (compare complete Table A.2).

In Table 5.2 (left-hand side) we can see the percentage of best performance in terms of word accuracy rate for each set of acoustic models tested on speech with noise N1 (subway) of the Aurora 2 evaluation set. On the right-hand side of the same table the relative confusion is shown indicating how often each set of models is selected by our approach for the test data from set N1 at different SNRs. A complete table also including all other noise domains N2 (babble), N3 (car), and N4 (exhibition hall) can be found in the Appendix in Table A.3. In the majority of cases the best performance is achieved when using the set of acoustic models from the same noise domain. But we can see that — especially for high SNRs — all other sets from the other noise conditions (but not from the clean conditions) perform equally well for a majority of utterances. One reason is the low complexity of the task of digit recognition. In combination with a rough quantisation of the word error rate due to often small numbers of words in an utterance, this often leads to exactly the same word error rates. But also the noise conditions seem to be rather similar and the way of speaking can even be considered to be equal and without Lombard effect, as noise is added artificially. This leads to much more similar features and acoustic models compared to the MoveOn evaluation set as can be seen in Figure 5.5. The figure depicts the centroids for a codebook

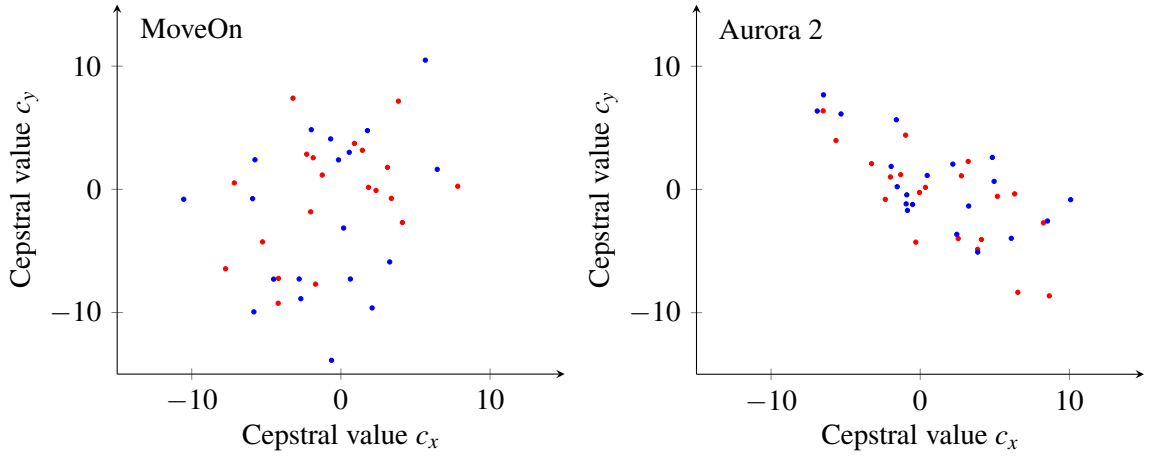


Figure 5.5: Comparison of cepstral variation for MoveOn and Aurora acoustic models. As an example the two most distant cepstral dimensions (c_x and c_y — measured by Euclidean distance) for the clean speech acoustic models of throat and close-talk microphone (MoveOn) and noise type N1 and N2 (Aurora 2) are plotted. The compact representation with $k = 20$ instead of the full set of acoustic models is used for easier comparison. The MoveOn acoustic models show a more distributed pattern, while both Aurora 2 acoustic models have a similar and more directed pattern.

of $K = 20$ showing two of their cepstral dimensions for clean close-talk and throat microphone acoustic models of the MoveOn evaluation set, on the left-hand side, and for the acoustic models of noise types N1 and N2 of the Aurora evaluation set, on the right-hand side. The MoveOn sets of acoustic models show a more spread pattern with larger distances compared to the Aurora sets of acoustic models with a much more similar distribution.

In general, our approach for blind acoustic model selection prefers the domain-specific acoustic models but not necessarily the best performing acoustic models for an utterance as we can also see in Table 5.3. Still, confusion seems to be more likely the lower the mismatch (the more similar the acoustic conditions) between training and test data is, so that the results with blind acoustic model selection are comparable to manually assigned domain-specific acoustic models.

	Domain	**Selected ...**	
test	**is best**	**is best**	**is in domain**
tc	70%	70%	100%
tn	83%	78%	91%
rc	85%	84%	96%
rn	93%	92%	98%
all	86%	84%	96%

Table 5.3: Selection rate for best set of acoustic models in terms of phoneme accuracy (MoveOn). The table provides the percentage of utterances per test set (tc:throat clean, tn: throat noisy, rc: right clean, rn: right noisy) for which the domain-specific and the automatically selected set of acoustic models perform best in terms of phoneme accuracy. In the last column the percentage of utterances for which the set of acoustic models from the same acoustic domain is selected is shown. The presented approach of acoustic model selection tends to classify by domain instead of best phoneme accuracy.

	is	**multi-cond.**		**bams**				
	domain	**16 GMs**	**64 GMs**	**ML,full**	**d_2,full**	**d_2,K=300**	**d_M,K=300**	**ideal**
tc	33.81	30.55	34.82	33.95	33.81	33.81	33.78	37.56
tn	52.87	47.44	53.85	52.71	52.32	52.07	51.93	54.93
rc	56.15	47.53	54.50	55.77	55.88	55.86	55.27	57.98
rn	53.98	48.63	54.89	53.89	53.65	53.62	53.53	54.86
all	51.48	45.96	52.12	51.36	51.12	51.03	50.87	53.18

Table 5.4: Phoneme accuracy rates in % for blind acoustic model selection (MoveOn). The first column presents the baseline results for a manual selection of one of the four domain-specific sets of acoustic models (tc:throat clean, tn: throat noisy, rc: right clean, rn: right noisy). Multi-conditional acoustic models with 16 and 64 Gaussian mixtures are evaluated. Blind acoustic model selection (bams) is tested for full sets of acoustic models as well as compact representations with $k = 300$. Different distance measures (ML: maximum likelihood, d_2: Euclidean distance, d_M: Mahalanobis distance) were evaluated. The theoretically possible performance (ideal) is compared in the last comlumn.

ASR Performance

The most relevant measure for our approach of blind acoustic model selection is the recognition accuracy of ASR. We compare the performance of our approach using different distance measures with a manual selection of domain-specific acoustic models and with the common approach of multi-conditional acoustic models. Table 5.4 shows the phoneme accuracies for the different experiments on the MoveOn evaluation set. The results for manually assigned domain-specific acoustic models is shown in the first column, the performance for multi-conditional acoustic models with 16 and with 64 Gaussian mixtures in the next columns. Blind acoustic model selection (*bams*) is tested for Euclidean distance d_2 and Mahalanobis distance d_M including the four domain-specific models. For comparison we also provide the results for the maximum-likelihood selection (ML) on the full set of acoustic models compared to Euclidean distance on the full set of models. The column *ideal* shows the theoretical best results for model selection when manually picking the highest phoneme accuracy of the four domain-specific models for each utterance. Neglecting the ideal case, we can see that except for the right clean data (*rc*) the multi-conditional acoustic models perform best increasing the average phoneme accuracy compared to the domain-specific models by about 0.5% absolute. The results of an ideal acoustic model selection (selecting from the four domain-specific models) would improve the phoneme accuracy in average by an additional 1% absolute compared to multi-conditional models. For throat clean (*tc*) and right clean data (*rc*), where only a small amount of training data is available, the improvement in phoneme accuracy is even more significant with 2 to 4% absolute improvement.

Unfortunately, in this setup these ideal values are not reached by our approach, which performs about equal to a manual selection of the domain-specific models. Mahalanobis distance and Euclidean distance provide comparable results. One reason for the gap between theoretically possible results (*ideal*) and our approach is that the initial approach tends to classify by domain (96% correctness) and not by best phoneme accuracy (84% correctness) as we discussed in the previous section. Also a classification method with stronger consideration of the probabilistic character of the speech recognition process, like the presented maximum-likelihood (ML) approach, shows only rather small improvements compared to Euclidean distance. Both, classification rate by domain as well as by best phoneme accuracy, increase only slightly to 98% and 86% when using ML.

The results on the Aurora 2 evaluation set show similar tendencies (Table 5.5). The multi-conditional acoustic models slightly improve the word accuracy rate by 1% absolute compared to manually selected

	is	**multi-cond.**		**bams**		
SNR	**domain**	**3 GMs**	**15 GMs**	d_2**,K=300**	d_M**,K=300**	**ideal**
clean	99.04	98.73	99.45	99.04	99.04	99.37
20dB	98.02	97.96	98.94	97.99	97.67	99.16
15dB	97.66	97.06	98.40	97.57	97.44	98.94
10dB	96.15	95.13	96.97	95.92	95.44	98.14
5dB	91.29	88.39	92.05	90.33	89.63	94.78
0dB	72.36	64.58	73.55	67.91	67.74	79.19
-5dB	32.39	26.68	34.65	28.10	26.06	43.38
all	83.84	81.22	84.86	82.41	81.86	87.57
20-0dB	91.10	88.63	91.98	89.94	89.58	94.04

Table 5.5: Average word accuracy rates in % for blind acoustic model selection (Aurora 2). The first column presents the averaged baseline results for a manual selection of one of the five domain-specific sets of acoustic models for different SNRs. Multi-conditional acoustic models with 3 and 15 Gaussian mixtures are evaluated. Blind acoustic model selection (bams) is tested for compact representations with $k = 300$ and different distance measures (d_2: Euclidean distance, d_M: Mahalanobis distance). The theoretically possible performance (ideal) is compared in the last column.

domain-specific acoustic models. Our initial approach with Euclidean distance provides slightly lower word accuracies (about 1% absolute) compared to a manual selection. Results mainly drop in lower SNRs but are comparable for SNRs of 10dB and above. Mahalanobis distance performs almost equal to Euclidean distance. An *ideal* selection of the best acoustic models in terms of best word accuracy rate would yield a clear improvement of about 2.7% absolute compared to the multi-conditional acoustic models with 15 Gaussian mixtures, especially for SNRs of 10dB and lower where absolute improvements reach up to 9%. This is remarkable, as, on the one hand, multiple models seem to be beneficial in case of low SNRs, but, on the other hand, classification by our approach for model selection suffers in case of such low SNRs.

As we discussed in the previous section for Table 5.1, noisy throat and close-talk microphone acoustic models often perform well on the clean test data of the same microphone and vice versa. A similar behaviour applies to the noisy models of the Aurora evaluation set independent of the noise type (Table 5.2). Thus, we decide to train only two domain-specific models for both evaluation sets joining clean and noisy training sets for each microphone of the MoveOn set as well as all noisy training sets of the Aurora evaluation set. We increase the number of mixtures of the trained acoustic models accordingly to 32 (MoveOn) and 12 (Aurora, noisy) to match the total number of previous mixtures and to cover the larger variability now modelled in each set of acoustic models.

The phoneme accuracies of our approach for blind acoustic model selection using a separate throat and close-talk microphone model are presented in Table 5.6. We now outperform multi-conditional acoustic models by 0.58% absolute using the full set of acoustic models for selection and still 0.5% in case of using a codebook of 300 centroids with a performance about equal to a manual selection. Also the ideal performance for blind acoustic model selection increases slightly. Only the clean test set of the close-talk microphone channel shows a degradation in ASR performance of about 2% absolute. The results for the Aurora evaluation set also improve compared to the setup of blind acoustic model selection with 5 different sets of acoustic models (Table 5.7). Here the results for a manual selection, multi-conditional acoustic models and blind acoustic model selection are about equal. The theoretical maximum (*ideal*) of blind acoustic model selection, on the other hand, drops by almost 3% absolute due to lower accuracy rates for low SNRs. This indicates that modeling variations by an increased number

	is	**multi**	**bams**		
	domain	**64 GMs**	d_2**,full**	d_2**,K=300**	**ideal**
tc	35.43	34.82	35.43	35.43	36.23
tn	55.37	53.85	55.32	55.32	55.73
rc	54.02	54.50	53.90	53.78	54.31
rn	55.34	54.89	55.17	55.03	55.83
all	52.80	52.12	52.70	52.62	53.27

Table 5.6: Phoneme accuracy rates in % for blind acoustic model selection, 2 AMs (MoveOn). The first columns present the baseline results for a manual selection of one of the two domain-specific sets of acoustic models and the multi-conditional acoustic models with 64 Gaussian mixtures for the four test sets (tc:throat clean, tn: throat noisy, rc: right clean, rn: right noisy). Blind acoustic model selection (bams) is tested for full sets of acoustic models as well as compact representations with $k = 300$ with Euclidean distance (d_2) in a setup with two domain-specific acoustic models. The theoretically possible performance (ideal) is again compared in the last column.

	is	**multi**	**bams**		
SNR	**domain**	**15 GMs**	d_2**,full**	d_2**,K=300**	**ideal**
clean	99.04	99.45	98.91	99.04	99.42
20dB	98.81	98.94	98.81	98.80	99.14
15dB	98.29	98.40	98.29	98.29	98.65
10dB	96.98	96.97	96.98	96.98	97.23
5dB	91.89	92.05	91.89	91.89	92.11
0dB	73.57	73.55	73.57	73.57	73.90
-5dB	34.66	34.65	34.66	34.66	36.58
all	84.75	84.86	84.73	84.75	84.92
20-0dB	91.91	91.98	91.91	91.91	92.21

Table 5.7: Average word accuracy rates in % for blind acoustic model selection, 2 AMs (Aurora 2). The first columns present the averaged baseline results for a manual selection of one of the two domain-specific sets of acoustic models and the multi-conditional acoustic models with 15 Gaussian mixtures for different SNRs. Blind acoustic model selection (bams) is tested for full sets of acoustic models as well as compact representations with $k = 300$ with Euclidean distance (d_2) in a setup with two domain-specific acoustic models. The theoretically possible performance (ideal) is again compared in the last column.

of Gaussian mixtures compared to separate acoustic models reduces the discrimination capabilities of the ASR system by an increased "blurriness" of each state of the HMMs.

Discussion

Our evaluations on the influence of the codebook showed that we are able to use a reduced set instead of the full set of mean vectors for blind acoustic model selection. We could reduce the MoveOn and the Aurora evaluation sets from 2400 and 546 to a codebook of about 100 centroids practically without changing the ASR performance of our approach of acoustic model selection. Our approach of blind acoustic model selection provided results close to a manual selection of the domain-specific acoustic models in all cases except for the five models setup of the Aurora 2 evaluation set, where we lost in performance compared to a manual selection. Reducing the number of specialised acoustic models to two improved the quality of the acoustic models — as the acoustic characteristics of the joined training sets are not too manifold — and provided sets of acoustic models, which significantly improved

	is domain	**multi 64 GMs**	**bams d_2,K=300**	**combined HMMs**	**concat. HMMs**
tc	33.81	34.82	33.81	33.69	29.80
tn	52.87	53.85	52.07	52.79	50.79
rc	56.15	54.50	55.86	55.46	53.03
rn	53.98	54.89	53.62	52.82	53.17
all	51.48	52.12	51.03	50.84	49.61

Table 5.8: Phoneme accuracy rates in % for model selection and model combination (MoveOn). The approaches of combining and concatenating domain-specific acoustic models are compared to a manual selection of one of the four domain-specific sets of acoustic models, the multi-conditional approach and blind acoustic model selection with four domain-specific sets of models. A combination of the domain acoustic models performs almost equal to blind acoustic model selection while the concatenation approach results in lower phoneme accuracy rates.

the recognition accuracy for acoustic model selection. For Aurora 2 we achieved results similar to multi-conditional acoustic models, while we outperformed such acoustic models in case of the MoveOn evaluation set. Multi-conditional acoustic models provide very good results for acoustic conditions, which are not to manifold, while blind acoustic model selection promises to outperform such acoustic models for application in various, rather different acoustic conditions.

5.5.3 Recombination of Acoustic Models

Our evaluations of our approach of blind acoustic model selection in case of 4 and 5 domain-specific models for the MoveOn and Aurora evaluation set showed a lower accuracy than multi-conditional acoustic models, while an ideal selection would yield improved results. Thus, we evaluate the effect of a recombination of the different sets of domain-specific acoustic models to analyse, if the maximum-likelihood decoding of the speech recogniser itself successfully selects the best performing acoustic models and provides accuracies closer to the ideal results of blind acoustic model selection. In Table 5.8 we compare the ASR performance of our approach of blind acoustic model selection with the two approaches of combining the separate sets of models into one general set of models as introduced in Section 5.3.1.

A combination of HMMs by concatenating the Gaussian mixtures and adapting weights and transition probabilities yield very similar results (-0.28% absolute) to our approach of blind acoustic model selection with separate acoustic models. A concatenation of the HMMs from the specific models also preserving the transition probabilities performs worse with about -1.5% absolute in phoneme accuracy. This indicates that one of the major benefits of using separate acoustic models is a specialisation of these models on their domain avoiding recognition errors caused by an increased “blurriness” of the classes to be recognised due to joined acoustic models including various acoustic conditions.

Discussion

A combination of the different sets of acoustic models neither by concatenation nor by combination of the HMMs yield any improvement compared to a multiple model approach with blind acoustic model selection. Even a small decrease in performance can be observed. Thus, depending on task and acoustic variations either a multi-conditional training approach or blind acoustic model selection with well adapted sets of acoustic models as compared in the previous section is recommended.

5.5.4 Relative Approach for Mismatch Compensation

Even though a reduction of acoustic information hardly influences the acoustic model selection process, using compact representation more likely influences the mismatch compensation step as additional mismatch is caused by the quantisation process. Thus, we exemplarily analyse the effect of the relative normalisation on the cepstral coefficients of reliable utterance pairs in a preliminary evaluation to find out about the general behaviour of the normalisation approach. Then we apply our relative normalisation approaches in case of full sets of acoustic models and on the codebooks of the same sets with a reduced set of reference vectors. That way, we see the possibly negative effect of the nearest neighbour assumption and the quantisation on the mismatch compensation.

Preliminary Evaluation

In a preliminary evaluation we analyse the behaviour of the two normalisation approaches in Section 5.4. Before we use the nearest neighbour information from the model selection step, we evaluate the performance for ideal conditions, adapting the following speech features to the following reference vectors:

1. MoveOn, throat microphone speech to close-talk microphone speech: As we have synchronous recordings, we have ideal reference vectors for each feature vector to adapt the features to. We have only microphone mismatch in the data.

2. Aurora 2, artificial noisy speech to clean speech: We have exactly the same utterances in clean speech and with artificial noisy speech for adaptation. We have only mismatch caused by additive noise.

3. TETRA Corpus, TETRA transmitted speech (TETRA radio) to clean speech (Clean 8): The utterances in the TETRA radio set are transmitted Clean 8 speech data. The only mismatch present is caused by the actual TETRA transmission.

In our setup we consider the aligned feature vectors of the second acoustic condition as reference set Y for normalisation and the feature vectors of the first condition as X, which should be normalised. We apply rQCN and rCGN accordingly on one example of each of the sets.

In Figures 5.6, 5.7, and 5.8 we see the original and adapted cepstral values for an example of each of the three sets. For rQCN and rCGN we see that the shape of the curve is not changed by these two approaches as only overall mean and gain are adapted. Generally, rQCN compared to rCGN results in a higher gain of the compensated feature values, as it normalises to a low/high percentage quantile (here $j = 4$) as opposed to rCGN normalising to a median gain. As we do not use multi-linear or non-linear transformations, the curve of the normalised cepstral coefficients are still more similar to original curve than the reference curve. Complex mismatch cannot be compensated and only in case of mismatch related to mean and gain improvements can be expected. The figures do not clearly show, if feature mismatch is reduced by our approach. Thus, an answer about the quality of the normalisation techniques in terms of ASR improvement will be provided in the following section by relevant recognition experiments in matched and mismatched conditions.

Compensation in Realistic Conditions

We empirically determined the weights $w_1 = 0.5$ and $w_2 = 0.5$ for rQCN and rCGN to be a good compromise between normalisation and introduced distortion. The threshold Θ for reliable pairs is

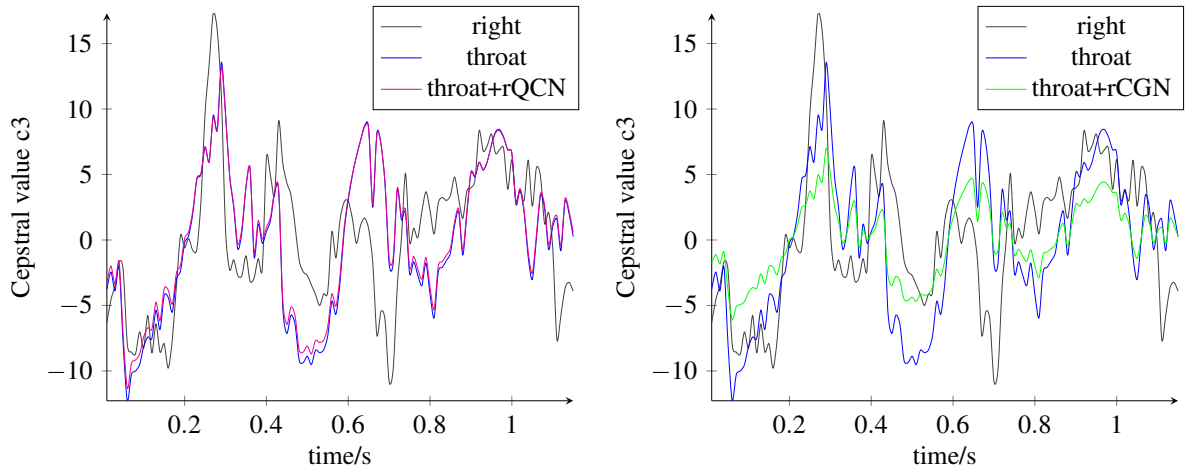


Figure 5.6: Cepstral feature normalisation with rQCN and rCGN (MoveOn). Relative cepstral normalisation with rQCN and rCGN is applied to the throat microphone speech features (throat) to reduce the mismatch to the close-talk microphone speech features (right) of the same utterance. The third cepstral value is shown as an example. While rQCN hardly changes the third cepstral feature of the throat microphone speech signal, rCGN changes mean and gain of the cepstral coefficients. The quite different shape of the throat microphone speech features is not compensated. A more complex mismatch compensation technique seems to be necessary to compensate for the major mismatch between these two channels.

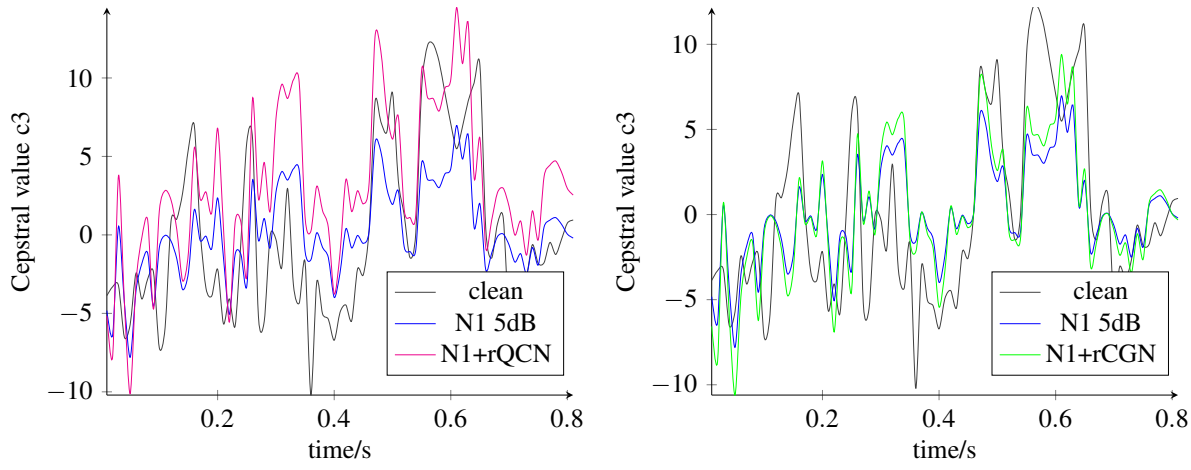


Figure 5.7: Cepstral feature normalisation with rQCN and rCGN (Aurora). Relative cepstral normalisation with rQCN and rCGN is applied to the simulated noisy speech features (N1 5dB) to reduce the mismatch to the clean speech features of the same utterance. The third cepstral value is shown as an example. While mean and gain are adapted, the shape of the curve is not affected. An obvious improvement in terms of reduced mismatch cannot be observed.

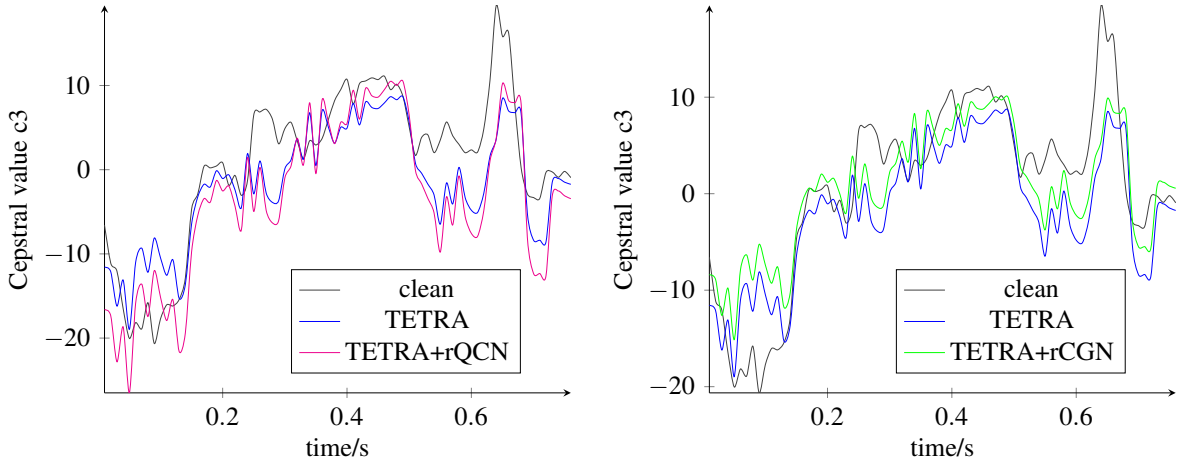


Figure 5.8: Cepstral feature normalisation with rQCN and rCGN (TETRA). Relative cepstral normalisation with rQCN and rCGN is applied to the TETRA transmitted speech features (TETRA) to reduce the mismatch to the clean speech features of the same utterance. Again, the third cepstral value is shown as an example. Mean and gain of the third cepstral coefficient are adapted. A significant mismatch reduction is not observed.

further defined by the average Euclidean distance $\Theta = 1/N \sum_{n=1}^{N} d_2(c_n^{(X)}, c_n^{(Y)})$ of all vector pairs. For rQCN we use the 4% quantile (rQCN_4 with $j = 4$) as in [70].

We consider different scenarios to evaluate the performance of our relative cepstral normalisation techniques. First, we analyse the case of no expected mismatch between the acoustic conditions of test and training data. We do not expect any significant increase in the ASR performance when applying our compensation approaches in such a case. Furthermore, the ASR performance should not decrease, which would be a sign for algorithmic mismatch introduced by our approaches. In Table 5.9 we can see the phoneme accuracy rates for the MoveOn evaluation sets for a manual domain selection and for blind acoustic model selection. For rQCN we get small improvements for both clean test sets (0.63 and 0.19% absolute), but also a small decrease in performance for the two other test sets (-0.44 and -0.63% absolute). As the clean speech test sets contain a lower number of utterances, the overall average phoneme accuracy on all utterances slightly decreases. The same effect can be seen for blind acoustic model selection. The change in performance has about the same magnitude for each set, which can be expected as the model selection performance is close to a manual selection. The approach of rCGN performs slightly worse than rQCN.

In in the appendix in Table A.6 providing the complete results for rQCN, we see for mismatched conditions that we gain improvements of about 0.65% absolute for both clean speech test sets in case of background noise but identical microphone channels. In particular for channel mismatch between throat and close-talk microphone the performance of the ASR performance slightly decreases for rQCN (between 0 and 0.69% absolute) in most cases. As mentioned in Section 2.4.1 mean and gain based methods mainly compensate additive noise, which reflects in a change of mean and variances of the features' statistics. Thus, complex channel effects caused by throat and close-talk microphones as analysed in detail in Section 4.3 are hardly compensated by these general statistics (compare Section 5.5.4) and we run the risk of introducing additional algorithmic distortion.

In Tables 5.10 and 5.11 we can find similar tendencies for both approaches in matched conditions and for blind acoustic model selection on the Aurora evaluation set. Furthermore, we have an insight on the influence of different SNRs on the results. For the case of no expected mismatch (manual selection and blind acoustic model selection in Table 5.10) the results of a word recognition are hardly affected by

	domain, no mismatch			bams, full		
	org.	**rQCN**	**rCGN**	**org.**	**rQCN**	**rCGN**
tc	33.81	34.44	34.00	33.81	34.44	34.00
tn	52.87	52.43	52.49	52.32	51.94	51.98
rc	56.15	56.34	55.96	55.88	56.10	55.74
rn	53.98	53.35	53.02	53.65	52.96	52.66
all	51.48	51.16	50.93	51.12	50.80	50.58

Table 5.9: Phoneme accuracy rates in % in matched conditions for rQCN and rCGN (MoveOn). Both approaches of relative normalisation are evaluated for a manual and automatic selection of the domain-specific and the best matching set of acoustic models. Small improvements by rQCN in matched conditions are achieved for throat clean (tc) and right clean (rc) conditions. In most other matched conditions recognition performance slightly decreases.

rQCN. The approach of rCGN does not significantly change the results in case of higher SNRs ($10dB$ and more), but leads to small to medium degradations (up to almost 4% absolute drop in word accuracy) for low SNRs.

In case of expected mismatch in Table 5.11 we can see the strength and weakness of the approach. Considering the clean speech acoustic models only, we gain an average of 0.7% and 0.51% absolute in word accuracy for rQCN and rCGN. For SNRs between $20dB$ and $0dB$ (as in the baseline evaluation measure in Section 3.1) rQCN even yields a 1.18% improvement with a maximum at $SNR = 10dB$ with up to 2.7% absolute improvement. This peak at $10dB$ can be explained by two factors influencing the rQCN normalisation approach: For SNRs above $10dB$ the mismatch to clean speech data is rather low and the effect of mismatch compensation cannot achieve significant improvements; for SNRs below $10dB$, on the other hand, the mismatch becomes larger but also the reference vectors determined by the nearest neighbour approach become less reliable. Thus, the risk of normalising towards mismatched quantiles increases neutralising the normalisation effect or even increasing the mismatch in the worst case. In case of general mismatch including all models for testing which do not belong to the domain of the test utterance (right-hand side of Table 5.11), no significant improvements can be reported (except for the clean speech test data with improvements of up to 0.58% absolute in word accuracy). We could already see in the performance and confusion table (Table 5.2) that the noise conditions seem to have rather similar characteristics as the recognition accuracies between the acoustic models trained on the the different types of noise do not vary as much as compared to the clean speech models. Furthermore, mean and gain as general statistics are influenced by all types of noise, so that any approach for normalisation that does not consider more specific noise characteristics, will only compensate a rather small part of more specific mismatch.

Discussion

We could identify mainly two issues influencing a relative mismatch compensation using estimated reference vectors: unreliable estimations of the reference vectors, and complex characteristics of the mismatch. In particular the usually mutual influence of both aspects prohibits more complex compensation techniques on an utterance level based on the proposed reference vectors.

As the approach of blind acoustic model selection aims at a small mismatch, mismatch compensation approaches in combination with the multi-model approach could not yield any improvements in the presented evaluations. In case of unavoidable mismatch caused by background noise, both approaches of relative cepstral normalisation — rCGN and especially rQCN — are able to improve the ASR results.

	domain, no mismatch			bams, full		
SNR	**org.**	**rQCN**	**rCGN**	**org.**	**rQCN**	**rCGN**
clean	99.04	99.06	99.02	99.04	99.06	99.02
20dB	98.02	97.97	98.02	97.98	97.95	98.03
15dB	97.66	97.70	97.66	97.59	97.69	97.62
10dB	96.15	96.10	96.02	95.93	95.88	95.81
5dB	91.29	91.10	90.45	90.39	90.17	89.52
0dB	72.36	72.18	68.46	68.31	67.85	64.51
-5dB	32.39	32.78	29.32	26.89	26.36	24.10
all	83.84	83.84	82.71	82.31	82.14	81.23
20-0dB	91.10	91.01	90.12	90.04	89.91	89.10

Table 5.10: Word accuracy rates in % in matched conditions for rQCN and rCGN (Aurora). Both approaches of relative normalisation are evaluated for a manual and automatic selection of the domain-specific and the best matching set of acoustic models. rQCN hardly affects the recognition performance in matched conditions while rCGN slightly reduces the word accuracy rate by about 1% absolute.

	clean models			all mismatched		
SNR	**org.**	**rQCN**	**rCGN**	**org.**	**rQCN**	**rCGN**
clean	99.04	99.06	99.02	86.82	87.36	87.40
20dB	93.51	93.29	93.80	96.10	96.05	96.17
15dB	80.59	81.75	81.60	91.99	92.22	92.28
10dB	57.11	59.81	58.44	83.53	84.01	83.82
5dB	31.13	33.19	31.38	70.15	70.11	69.85
0dB	11.97	12.19	12.30	46.14	45.29	44.69
-5dB	4.86	3.84	5.23	17.62	16.41	16.87
all	54.03	54.73	54.54	70.34	70.21	70.15
20-0dB	54.86	56.04	55.50	77.58	77.54	77.36

Table 5.11: Word accuracy rates in % in mismatched conditions for rQCN and rCGN (Aurora). Both approaches of relative normalisation are evaluated for a typically case of mismatched conditions (evaluation on clean speech acoustic models) as well as for all mismatched combinations. In all mismatched conditions rQCN and rCGN perform about equally well as the unmodified features (org.). In case of clean speech acoustic models a small average improvement can be seen for both approaches with clear improvements ($> 2\%$ absolute) for medium SNRs around 10dB.

In such expected situations the approaches for relative normalisation can also be applied without blind acoustic model selection.

5.5.5 Evaluation on LVCSR

In the previous sections we evaluated our approach of acoustic model selection and our approaches for relative cepstral normalisation. In case of an appropriate configuration experiments showed good results and often small improvements compared to commonly used alternative setups. Still, the evaluation was limited to two small datasets with comparatively simple tasks of ASR based on full word HMMs (Aurora 2) and monophone HMMs (MoveOn). While these setups comprise acoustic models with about 600 and 2400 Gaussian distributions, our LVCSR setup with the TETRA Corpus uses triphone HMMs with an overall of about 200,000 Gaussian distributions in the full set of acoustic models. The use of the compact representation instead of the full set of mean vectors for each acoustic models set is much more relevant in such a use case.

In the following paragraphs we briefly evaluate our approach of acoustic model selection for a setup with two sets of acoustic models. The setup includes the Clean 8 acoustic models and the TETRA Radio acoustic models with the respective test sets. Thus, we evaluate the case of rather different channel characteristics with TETRA transmitted speech data, on the one hand, and clean speech TV broadcast data, on the other hand. Again, we use a compact representation provided by a codebook of 300 vectors, which equals a reduction of information compared to a full set of mean vectors by about 1:650. Furthermore, we test the approaches for relative cepstral normalisation (rCGN and rQCN) on an extended setup also including AMR 4.75 and TETRA Codec acoustic models while using the TETRA Radio test set for mismatched conditions.

Acoustic Model Selection

In Table 5.12 the results for blind acoustic model selection are listed. As opposed to the two-models setup for Aurora and MoveOn sets, we lose a bit of ASR accuracy in case of TETRA Radio test data compared to a manual selection (+2.8% absolute WER) when using our approach with a compact representation of size $K = 300$. The Clean 8 test data is not affected and provides equal results to a manual selection. Our model selection approach selects the acoustic models from the same domain in about 100% (more precise 99.983%) for the Clean 8 test data — practically equal to the manual selection — and in about 90% for the TETRA Radio test data. Similar to the Aurora and MoveOn evaluations we can see that an ideal selection would reduce the word error rate by about 3.1% absolute compared to multi-conditional acoustic models. In practice our approach in average performs equally well to multi-conditional acoustic models providing a better performance on Clean 8 data (-1.3% WER absolute) and a worse performance on TETRA Radio data ($+1.4\%$ WER absolute).

Mismatch Compensation

For evaluation of our approaches for relative cepstral normalisation (rCGN and rQCN) we employ an evaluation setup providing mismatch between training and test data. Therefore, we use the TETRA Radio test data for evaluation on several acoustic models providing certain mismatch. The sets of acoustic models are trained on Clean 8, AMR 4.75 and TETRA Codec training data. Each utterance of the test set is normalised with rCGN and with rQCN relative to the compact representation with $K = 300$ of each set of acoustic models. The same parameters as in Section 5.5.4 are used for rCGN and rQCN.

Table 5.13 shows the performance of TETRA Radio test data on the mismatched acoustic models with and without normalisation. The baseline results are poor with a word error rate up to 63.4% in

test	Single model domain	Single model multi	bams k=300	bams ideal
Clean 8	29.1	30.4	29.1	27.5
TETRA Radio	42.3	43.7	45.1	40.5
Mean	35.7	37.1	37.1	34.0

Table 5.12: WER in % for blind acoustic model selection on TETRA evaluation set. The WER for blind acoustic model selection with a compact representation of the acoustic models with $k = 300$ and Euclidean distance is compared to the results for a manual selection of domain-specific sets of acoustic models and of multi-conditional acoustic models. The theoretically possible (ideal) results in the last column complete the table.

training	HTK	rCGN	rQCN
Clean 8	63.4	62.9	62.5
AMR 4.75	62.5	62.6	62.6
TETRA Codec	62.8	62.8	62.9

Table 5.13: WER in % for rQCN and rCGN in mismatched conditions of the TETRA evaluation set. The TETRA Radio test set is evaluated on mismatched acoustic models trained on Clean 8, AMR 4.75 and TETRA Codec training data. The baseline results (HTK) are slightly improved by 0.5% and 0.9% absolute with rCGN and with rQCN in case of the Clean 8 acoustic models. No improvements are achieved for the acoustic models base on the AMR and TETRA Codec data.

case of Clean 8 models caused by the mismatch between training and test data. In case of the Clean 8 acoustic models we are able to slightly improve the results by 0.5% and 0.9% absolute applying rCGN and rQCN. In particular the aspect of relative mean normalisation should have a positive effect on the features, as we mainly expect channel mismatch to be present. Applying relative normalisation on the setup with acoustic models from AMR 4.75 and TETRA Codec training data does not provide any improvements. As either AMR 4.75 Codec or the similar TETRA Codec are also used in the actual TETRA transmission via the TETRA devices, any channel mismatch between data affected by these two codecs and data transmitted via the TETRA radio channel is presumably lower than between Clean 8 data and TETRA Radio data. Additive noise is practically not present. Thus, the two major sources of mismatch targeted by these normalisation techniques are hardly present, and hence, the recognition results do not improve in the latter case.

Discussion

Also for LVCSR acoustic model selection provides results close to a manual selection. Still the complexity of the acoustic models and the high compression to a much smaller codebook slightly increase the WER. Compared to multi-conditional acoustic models blind acoustic model selection performs about equally well. The theoretical case of an ideal model selection again improves the recognition results by several percent absolute. All in all the observed tendencies are very similar to the evaluations on the MoveOn and Aurora 2 evaluation sets. The two relative normalisation techniques rCGN and rQCN are capable of reducing some of the linear mismatch in the expected cases even though we have to deal with presumably less precise reference vectors due to a larger influence of quantisation effects caused by using the codebook representation of the acoustic models. Our experiments indicate that our approaches are not limited to rather simple ASR tasks and setups but could also be useful in much more complex LVCSR tasks. Still, particular attention must be paid in this case due to the higher complexity of the acoustic modelling for LVCSR tasks. Evaluations on more realistic scenarios with realistic noisy speech

from acoustically different domains are necessary for a better insight on blind acoustic model selection and LVCSR.

5.6 Conclusion

Motivated by the previous chapters we introduced and evaluated a new concept for blind acoustic model selection. This concept makes use of the shortcomings of commonly used acoustic features incorporating unwanted but significant information from speech variabilities and the acoustic environment. We argued, why this information is probably sufficient for a blind selection of the acoustic models, and why this information might even be reduced without introducing significant errors in the recognition process. We presented an approach for reducing the sets of acoustic models to a compact codebook as well as a selection approach based on minimum distance classification. The reduction to a codebook of an appropriate size did not show any significant decrease in the ASR performance for the acoustic model selection approach. The evaluation of our initial acoustic model selection approach resulted in a good overall performance with similar results to a manual selection of the acoustic models from the same domain. Still, the performance of multi-conditional acoustic models were not always reached, even though the theoretically possible performance of the presented concept of acoustic model selection would generally outperform even multi-conditional acoustic models. This is mainly caused by the drawback of our approach that it tends to classify by acoustic domain and not by best performance in terms of ASR accuracy. With a careful selection and training of a smaller number of different sets of acoustic models with more different acoustic characteristics, we were able to reach or even improve the performance of the multi-conditional acoustic models.

We could further identify some advantages compared to multi-conditional acoustic models and many other approaches: Our approach works completely blindly without assuming any source of distortion or speech variability still achieving results comparable to a manual selection in most situations. Thus, arbitrary acoustic models just based on the same types of features can be easily added to the system to extend the coverage of acoustic conditions with good recognition performances. Neither retraining nor adaptation of acoustic models is necessary, so that this approach can also be considered, when no training data but only the acoustic models are available. When using a compact representation of the sets of acoustic models the additional processing time is generally low even for complex triphone models. Even though this aspect was not evaluated here, we believe that the selection process is also capable of differentiating between different languages in case of multi-lingual speech recognition as different languages provide variations in typical speech sounds and as our approach works even for acoustic models based on different phoneme sets. This specific aspect will be evaluated in future work.

In general, our evaluations showed that multi-conditional acoustic models do not necessarily provide the best results as multiple acoustic models with an ideal selection would outperform this approach. One reason is probably the even weighting for all utterances assuming a representative distribution of training data over all possible conditions, which is often not the case. Another aspect is the increased risk of misclassification of phonemes (substitutions) due to higher variances of the Gaussian distributions and an increased number of mixtures per state. Our results indicate that either a consideration of uneven distributions over the various acoustic conditions must be taken into account for training improved multi-conditional acoustic models or the selection process for multiple acoustic models must be improved to achieve optimal results even in situations, where we did not reach the performance of multi-conditional acoustic models yet.

In case of unavoidable acoustic mismatch we also presented two relative normalisation approaches, relative cepstral gain normalisation (rCGN) and relative quantile based cepstral dynamics normalisation

(rQCN), incorporating information from the acoustic model selection step to enable an online feature normalisation. While both approaches showed good results in several cases of expected linear mismatch, for example, between clean and noisy speech, they failed to improve the results in other scenarios, where more complex cepstral mismatch is expected. Such mismatch includes speech and feature variabilities, which are often non-linear and cannot be compensated by adapting general statistics, and thus, are out of reach of both normalisation techniques. A general risk of such approaches is the assumption that the information from the model selection step provide information reliable enough for a relative normalisation. If this is not the case, compensation effects might be equalised by an erroneous estimation of reference vectors. In particular for non-linear compensation methods, which we did not consider here, this assumption would be even more critical, as adaptation to unreliable reference vectors can introduce more significant distortion to the features. In the evaluated case of the linear normalisation techniques of general statistics we could show that we were able to further improve the recognition results in several mismatched conditions even though cepstral mean normalisation was already applied to the speech features before.

Chapter 6

Conclusion

In this work we evaluated and summarised various aspects relevant in robust ASR. In the fundamentals section we motivated the state-of-the-art in ASR and discussed one of the major fields of research in this area: the problem of acoustic distortion and speech variabilities in ASR as well as methods for their compensation. With the aim of investigating in actual problems of real-life scenarios and the lack of available corpora suited for our evaluations, we developed two corpora capable of analysing relevant effects caused by different sources of distortion. In particular the MoveOn Corpus with realistic noisy speech and two rather different microphone channels proved to be valuable in research in this area. For that reason we decided to make this corpus available to the scientific community.

For a better understanding of the problems of ASR degradation caused by acoustic distortion, we separately evaluated the effects of background noise, microphone channel, TETRA hardware, coding/decoding algorithms and low-pass filtering on simulated and realistic data enabled by the two corpora created and presented before. We pointed out that even small distortion can cause rather significant changes in the speech features leading to a poor ASR performance. Furthermore, we showed that a simulation of the distortion is rather difficult and hardly successful comparing the resulting signal and speech features of the simulation with the realistic data. This was accounted to the many possible influences on realistic data and the high variability of changes in the features caused by these influences. The complexity and non-linearity of certain types of distortion often influencing different frequency regions and phonemes in a different way make a simulation as well as a compensation difficult. With our experiments we were able to emphasise the often stated notion that current speech features (like MFCCs used in this work) are very vulnerable to any acoustic variation and are not ideal for robust ASR.

One way of improving the situation is the implementation of improved speech features less affected by the acoustic environment. Unfortunately, no major improvements were reported in this direction in the last decades. Another way is the training or adaptation of multi-conditional or multiple acoustic models to arrange with this problem by trying to cover as much of the variability in speech and acoustics as possible to have a chance to perform well in ASR in many different situations. This second option motivated a new approach of acoustic model selection presented in this work. Our multi-model approach is called blind acoustic model selection, as we use the speech features and the acoustic models directly to find the best matching set of acoustic models for a given utterance without any additional assumption about speech characteristics or type of distortion. This approach considers the results of the previous evaluations that indicated that the amount of information about the acoustic conditions captured by the features is rather significant. We showed that multiple acoustic models can indeed perform better than multi-conditional acoustic models, if an ideal model selection reliably selected the best performing acoustic models from a set of multiple models. In practice our presented approach achieves good results similar to a manual selection. For rather diverse acoustic conditions and a careful selection of the sets of acoustic models our approach was also able to improve the speech recognition accuracy compared to multi-conditional acoustic models.We believe that our concept of blind acoustic model selection provides an alternative direction for research and system development in case of situations, where an

ASR system has to deal flexibly with various, rather diverse acoustic conditions. Even though the initial approach did not always outperform multi-conditional acoustic models, we think that further research in this direction eventually enables comparable or superior results in many more situations, as the theoretical performance of multiple models indicated.

We stressed in this work that a robust ASR system should aim at avoiding mismatch instead of compensating mismatch whenever possible. Still, we need to consider mismatch compensation in reality, as smaller mismatch can hardly be avoided due to the sensitivity of the speech features and the variability of speech. Thus, we further evaluated relative mismatch reduction approaches. We considered two common, successfully applied approaches of mean and gain related feature normalisation to enable a post-normalisation relative to the estimated reference vectors based on our approach of blind acoustic model selection. As complex mismatch cannot be compensated by the selected normalisation methods our approaches did not provide clear improvements on all evaluation data. But in all cases where mean and gain of the signal are expected to be affected by the type of mismatch, we were able to provide improvements of up to several percent absolute in recognition accuracy even though a cepstral mean normalisation was already applied before.

Future work in robust ASR can benefit from the knowledge about mismatch and acoustic information in speech features gained in our experiments. In particular our approach of blind acoustic model selection offers various directions of future research. Our initial approach showed a tendency to classify by acoustic domain (similar to a manual selection) but did not optimally consider the best performing acoustic models in terms of ASR accuracy. Here, different classification methods or distance measures might lead to improvements and to a performance closer to the theoretically possible results. Other directions of future research might include improved training processes of the acoustic models already considering a blind acoustic model selection during ASR. One promising way is an improved automatic clustering of the utterances into clusters of similar acoustic conditions within each cluster and rather different conditions beyond the cluster. Especially the improved results when joining office (clean) and motorcycle (noisy) training sets into one training set indicate that the preparation of appropriate acoustic models is a key issue for blind acoustic model selection. Further improvements might also be possible in the area of mismatch compensation based on the selection process, if the problem of unreliable feature vectors can be tackled successfully. While we focused on statistical features for normalisation, which are less affected by unreliable reference vectors, compensation of more complex mismatch requires multi-linear or non-linear transformations. Such methods will be more sensitive to outliers in the reference vectors. One way to enable such approaches and also to improve the presented normalisation techniques could be an estimation of a reliability value for each of the reference vectors with a better consideration of this reliability during the normalisation process.

As long as we do not find much better features for ASR, we have to cope with the insufficiencies of the current features. Until such better features are available, we suggest to consider to use the weakness of these features (i.e. the inclusion of a considerable amount of undesired acoustic information) in future scientific work for classification and compensation of these variabilities as proposed in this work.

Chapter 7

Scientific Achievements

In this work we provided a detailed overview and analysis of challenges for automatic speech recognition (ASR) under non-ideal acoustic conditions. We evaluated various aspects of acoustic distortions and their complex influence on speech features commonly used for ASR. These extensive evaluations were enabled by two speech corpora that we purposely design for evaluating acoustic distortion and robust ASR approaches. They provided several different sources of acoustic distortion from background noise and microphone channel characteristics to hardware and coding effects, with the possibility of a separate evaluation of one or the other influence. In a final chapter we further introduced a new approach for blind acoustic model selection that was motivated by the results of the previous chapters. An evaluation of our approach proved its usefulness for ASR in situations of expected diverse acoustic conditions showing a performance similar to a manual selection of the acoustic models. It showed to be capable of improving even multi-conditional acoustic models in situations of various acoustic conditions and a careful definition of the training sets for each of the multiple models. Two new relative cepstral normalisation techniques were introduced, which provided improved speech recognition accuracies in several mismatched conditions even if cepstral mean normalisation was already applied before. It further enables a post normalisation in case of a feature extraction and acoustic model training setup that did not consider normalisation beforehand.

In more detail the major scientific achievements in this work include the following:

- We developed a unique speech and noise corpus, the MoveOn Motorcycle Speech and Noise Corpus, enabling intensive research of various sources of distortion. This corpus proved to be helpful in research of robust ASR in several ways, as we demonstrated in the subsequent chapters, where we intensively used the MoveOn Corpus due to some of its unique characteristics. This includes the possibility to separate and evaluate certain effects of acoustic distortion on ASR. The MoveOn Corpus is planned to be available to the scientific community from end of 2012 for evaluations on noisy speech and robust ASR.

- We provided evaluations on major types of acoustic distortion and their impact on ASR in depth and extent beyond any scientific work known to us. Our results offered deepened insights on feature mismatch and influences on ASR caused by these distortions. The evaluations included a detailed analysis of the following major sources relevant for ASR:
  - Background noise: In an extensive and detailed evaluation on the influences of environmental noise and differences between realistic and simulated additive noise, we identified several problems present in common assumptions used for noise simulation and noise reduction. We could show that background noise indeed cannot be considered additive as often assumed. While several scientific publications already indicate that this assumption is not necessarily true, we could show that simulated noise based on the additive assumption is neither very similar in terms of ASR results nor in the spectral and cepstral characteristics.

We could identify some similarities but also major differences even in our ideal conditions for simulation. One of the challenges for simulation is the Lombard effect, which is highly variable and would require a much more complex mathematical model for noise simulation to improve the simulation quality. To our knowledge this is the first evaluation that clearly showed these aspects in the context of ASR.

- Microphone channel: Due to synchronous recording of two rather different microphone channels in the MoveOn Corpus — a throat microphone and a close-talk microphone — we were able to evaluate and identify various differences in both channels. Some aspects already evaluated in other scientific work could generally be confirmed by our experiments, while we could also identify and discuss further effects on the speech recognition process caused by the microphone channel differences. The different influences of both channels on the speech signal and the cepstral features are complex and strongly dependent on the type of phoneme and the frequency, which makes a compensation of mismatch between these channels very difficult.
- Channel effects: In an evaluation of the TETRA radio channel we analysed step by step the effect of (transmission) channel distortions including hardware and coding effects. This is to our knowledge the most extensive evaluation of the TETRA channel for ASR done so far. Based on the TETRA channel with low-pass filtering effects, coding/decoding, and hardware influences we identified certain severe effects on common speech features and ASR performance, which can also be expected for other transmission channels or hardware setups. Especially the introduction of harmonics by the hardware and their amplification by the TETRA coding caused serious distortion and degradation of the ASR performance.

- We proposed and evaluated a new approach for multi-model speech recognition in situations with diverse sources of distortion. This approach of blind acoustic model selection for ASR was motivated by the previous results, which consolidated the assumption that a universal approach for mismatch compensation is hardly possible, and the best way of enabling ASR in distorted environments is a proper adaptation to the acoustic domain. Thus, we proposed a concept of acoustic model selection not making any assumptions about the type of distortion. To our knowledge such an unrestricted approach for multi-model ASR directly using the speech features and the acoustic models for classification is new. Our approach enables flexible ASR systems easily adaptable to new acoustic conditions, whenever sufficient adaptation data or a set of acoustic models for these conditions is available. The performance is similar to a manual selection and — dependent on the acoustic conditions and the setup of the system — capable of providing improved recognition results compared to multi-conditional training.

- We introduced two new approaches for a relative cepstral feature normalisation. With these approaches we investigated whether the estimation of a reference feature vector for normalisation by our nearest neighbour approach from acoustic model selection also provides sufficient information for a relative normalisation of extracted speech features. The extension of typical normalisation techniques aims at a normalisation of mean and gain related characteristics of the speech features and do not compensate more complex acoustic mismatch. Both approaches provide improved recognition accuracies in mismatched situations of up to a several percent absolute, whenever mismatch affecting mean and gain of the features can be expected. More complex mismatch cannot be compensated by these methods and would need different approaches.

We believe that we accounted for relevant and valuable information and insights into the area of acoustic distortion and feature mismatch caused by various sources of distortion. Our approach for

blind acoustic model selection identified a new direction for multi-model speech recognition with several advantages compared to multi-conditional acoustic models and even improved performances in terms of ASR accuracy in several setups. Our suggested approaches of relative cepstral normalisation derived from the blind acoustic model selection algorithm enable a post-normalisation even if acoustic models did not incorporate a feature normalisation step beforehand.

Appendix A

Additional Information and Results

This chapter of additional information and results contain more details and extended tables and figures helpful for understanding several aspects and evaluation results introduced and discussed in the main part of this thesis. While we believe that this information is relevant, we tried to concentrate on the central information and results in the previous chapters. Thus, we decided to present this additional information at the end to avoid too many interruptions while reading this work. Tables and figures of the appendix are usually linked within the text at the related parts of the work.

Phoneme	Example	Transcript	Phoneme	Example	Transcript
p	pin	p I n	w	wasp	w Q s p
b	bin	b I n	j	yacht	j Q t
t	tin	t I n	I	pit	p I t
d	din	d I n	e	pet	p e t
k	kin	k I n	{	pat	p { t
g	give	g I v	Q	pot	p Q t
tS	chin	tS I n	V	cut	k V t
dZ	gin	dZ I n	U	put	p U t
f	fin	f I n	@	another	@ n V D @
v	vim	v I m	i:	ease	i: z
T	thin	T I n	eI	raise	r eI z
D	this	D I s	aI	rise	r aI z
s	sin	s I n	OI	noise	n OI z
z	zing	z I N	u:	lose	l u: z
S	shin	S I n	@U	nose	n @U z
Z	measure	m e Z @	aU	rouse	r aU z
h	hit	h I t	3:	furs	f 3: z
m	mock	m Q k	A:	stars	s t A: z
n	knock	n Q k	O:	cause	k O: z
N	thing	T I N	I@	fears	f I@ z
r	wrong	r Q N	e@	stairs	s t e@ z
l	long	l Q N	U@	cures	k j U@ z

Table A.1: SAMPA British English phoneme set. The table lists all SAMPA phonemes grouped by phoneme type as used in the MoveOn Corpus. For each phoneme a word example including SAMPA transcript is given. The examples are taken from http://www.phon.ucl.ac.uk/home/sampa/english.htm.

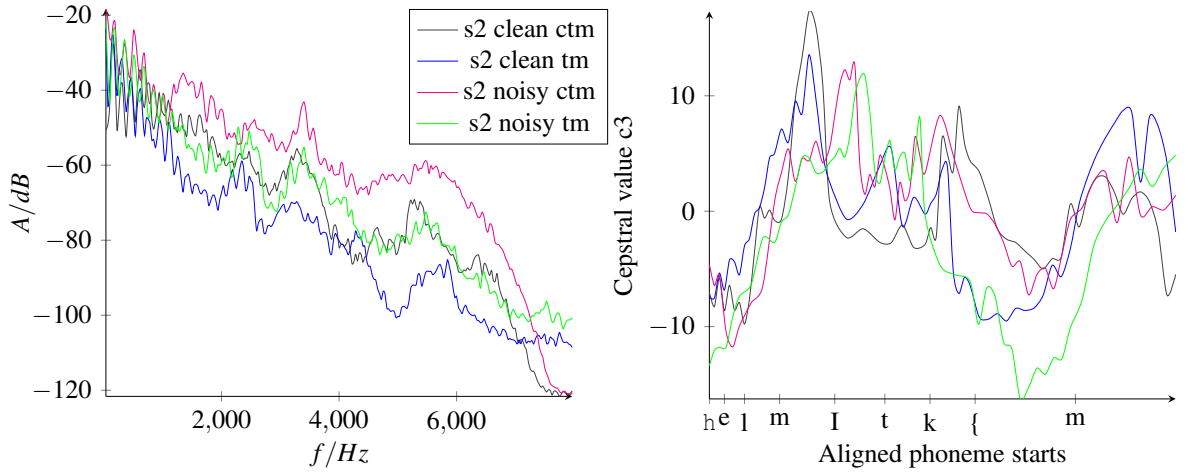


Figure A.1: Example for influences on spectrum and cepstral values caused by channel. The figure on the left hand side shows the spectrum for the utterance "Helmet Cam" for speaker *s*2 from Figure 4.9 for clean and noisy conditions recorded with throat (*tm*) and close-talk microphones (*ctm*). The figure on the right hand side compares the aligned third cepstral coefficients for the same examples. Various effects of channel characteristics and background noise on the speech signal are demonstrated.

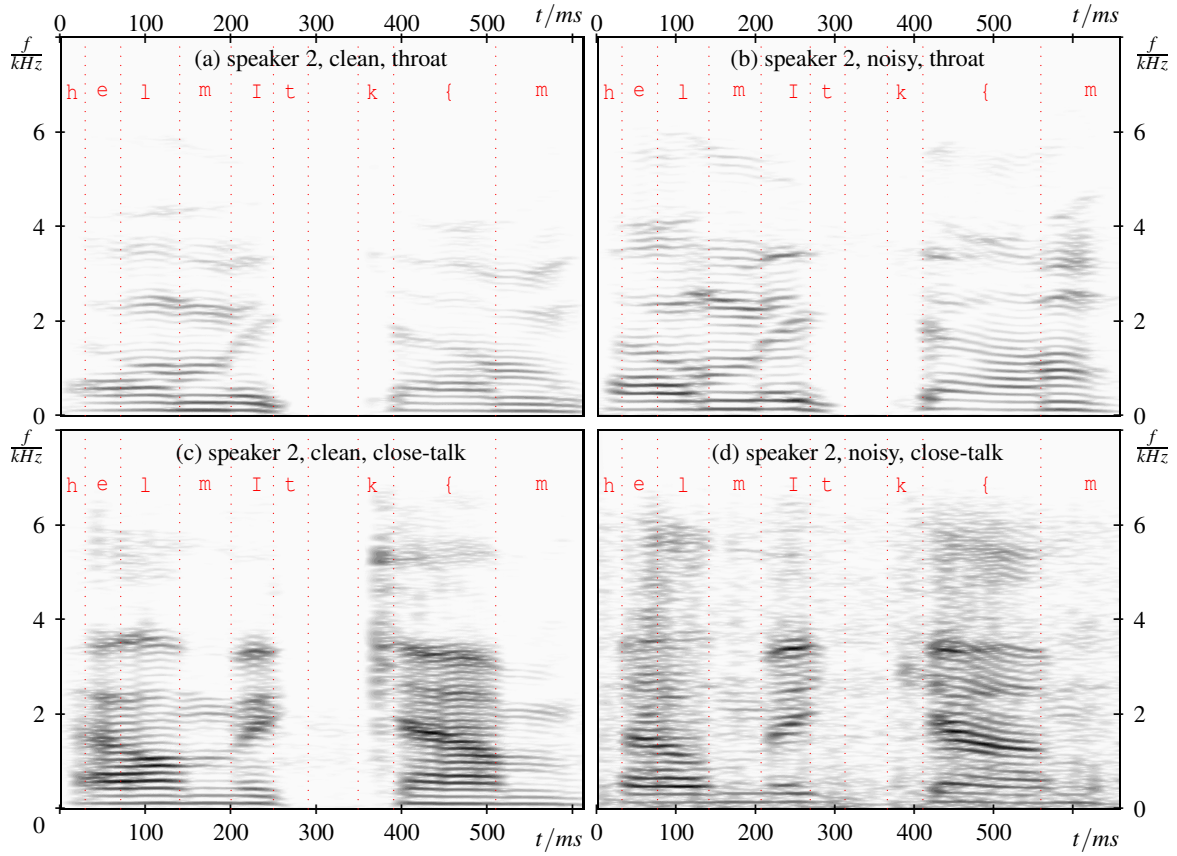


Figure A.2: Spectrograms for channel and acoustic variations of same speaker. The spectrograms show the utterance "Helmet Cam". (a) and (c) as well as (b) and (d) are synchronous recordings of the throat and close-talk microphone of the same utterance and the same speaker in clean and in noisy acoustic conditions. Next to the effects visible in Figure 4.10 also a lower amount of background noise in the throat microphone signal as well as changes in the speaking characteristics in noisy compared to clean environments can be observed.

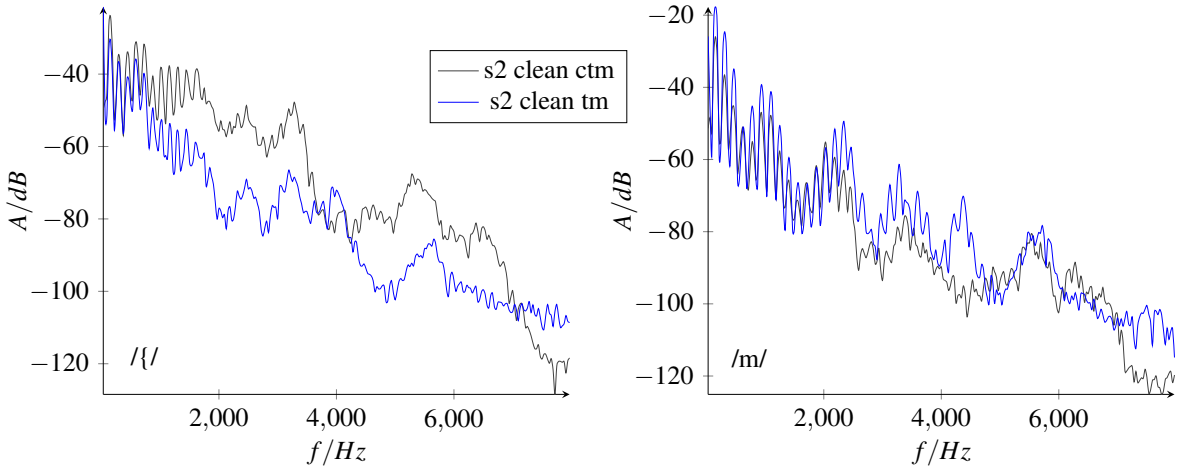


Figure A.3: Influence of the channels on the spectrum of the phonemes /{/ and /m/. The spectra for both phonemes show various differences for throat microphone (*tm*) and close-talk microphone (*ctm*) channel for synchronous recordings of speaker *s*2 in clean conditions. Differences include frequency dependent differences in the peaks' amplitudes as well as peaks missing in one of the signals while being clearly visible in the other one. (Identical to Figure 4.12 to enable easy comparison with Figure A.4 here.)

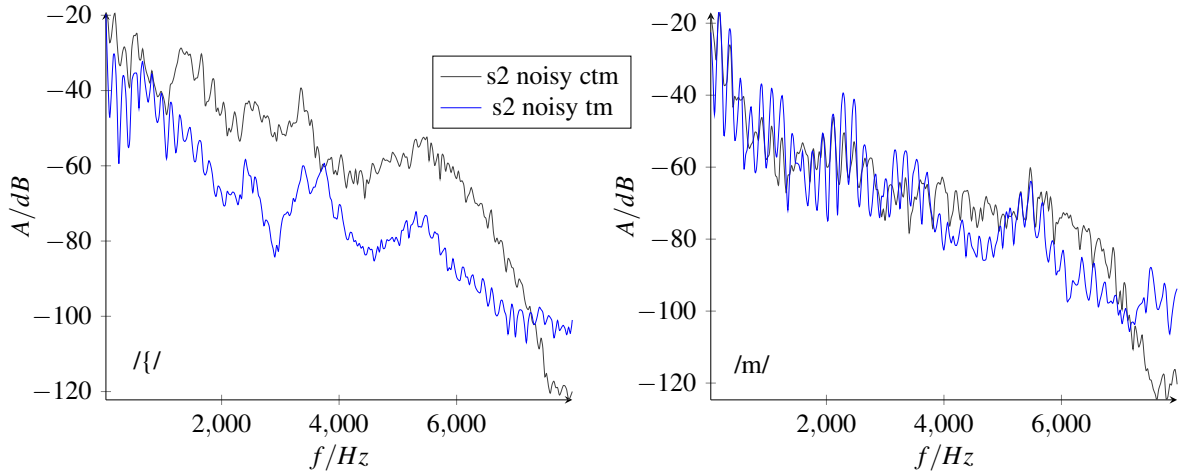


Figure A.4: Influence of the channels on the spectrum of the phonemes /{/ and /m/. The spectra for both phonemes show various differences for throat microphone (*tm*) and close-talk microphone (*ctm*) channel for synchronous recordings of speaker *s*2 in noisy conditions. Compared to Figure A.3 with clean speech, the spectrum of the phonemes is affected by background noise reducing characteristic peaks in particular for the close-talk microphone channel. The throat microphone channel is less affected.

	Acoustic models					**Acoustic models**			
is best	tc	tn	rc	rn	**full**	tc	tn	rc	rn
tc	70%	37%	17%	4%	tc	0.99	0.01	0.00	0.00
tn	28%	83%	5%	6%	tn	0.08	0.91	0.00	0.00
rc	5%	4%	85%	21%	rc	0.00	0.01	0.97	0.02
rn	3%	11%	8%	93%	rn	0.00	0.01	0.02	0.97
k=500	tc	tn	rc	rn	**k=300**	tc	tn	rc	rn
tc	0.99	0.01	0.00	0.00	tc	1.00	0.00	0.00	0.00
tn	0.10	0.89	0.00	0.01	tn	0.09	0.91	0.00	0.01
rc	0.01	0.01	0.95	0.03	rc	0.01	0.01	0.96	0.03
rn	0.00	0.01	0.01	0.98	rn	0.00	0.01	0.01	0.98
k=100	tc	tn	rc	rn	**k=50**	tc	tn	rc	rn
tc	0.99	0.01	0.00	0.00	tc	0.99	0.01	0.00	0.00
tn	0.11	0.88	0.00	0.01	tn	0.09	0.90	0.00	0.01
rc	0.00	0.02	0.94	0.05	rc	0.02	0.01	0.92	0.04
rn	0.00	0.01	0.01	0.98	rn	0.00	0.01	0.01	0.98
k=10	tc	tn	rc	rn	**k=5**	tc	tn	rc	rn
tc	1.00	0.00	0.00	0.00	tc	0.99	0.01	0.00	0.00
tn	0.11	0.86	0.00	0.03	tn	0.21	0.74	0.00	0.05
rc	0.07	0.01	0.81	0.11	rc	0.08	0.04	0.72	0.16
rn	0.00	0.04	0.01	0.95	rn	0.00	0.05	0.03	0.92

Table A.2: Full confusion matrix for blind acoustic model selection (MoveOn). Upper left hand side of the table shows the percentage of utterances of each test set (tc: throat clean, tn: throat noisy, rc: right clean, rn: right noisy) for which the set of acoustic models provide the best phoneme accuracy rates. Acoustic models without mismatch (diagonal) usually perform best. The remainder of the table presents the confusion matrix for blind acoustic model selection and compact representations of different sizes k. Confusion slightly increases for compact representation of a low dimension k.

	Acoustic models (is best)						Acoustic models (confusion)				
N1, best	clean	N1	N2	N3	N4	**N1, 300**	clean	N1	N2	N3	N4
clean	99%	64%	81%	77%	49%	clean	1.00	0.00	0.00	0.00	0.00
20 dB	86%	97%	95%	94%	96%	20 dB	0.01	0.84	0.03	0.01	0.11
15 dB	64%	96%	91%	89%	94%	15 dB	0.00	0.88	0.02	0.00	0.10
10 dB	27%	94%	84%	83%	91%	10 dB	0.00	0.89	0.03	0.02	0.06
5 dB	9%	89%	70%	68%	84%	5 dB	0.00	0.77	0.07	0.04	0.12
0 dB	6%	80%	50%	56%	70%	0 dB	0.00	0.62	0.22	0.12	0.03
-5 dB	19%	66%	43%	56%	62%	-5 dB	0.00	0.21	0.16	0.58	0.05
N2, best	clean	N1	N2	N3	N4	**N2, 300**	clean	N1	N2	N3	N4
clean	99%	64%	79%	76%	49%	clean	1.00	0.00	0.00	0.00	0.00
20 dB	79%	93%	98%	92%	93%	20 dB	0.03	0.03	0.87	0.04	0.03
15 dB	48%	90%	96%	87%	91%	15 dB	0.00	0.02	0.92	0.04	0.02
10 dB	18%	81%	94%	76%	79%	10 dB	0.00	0.01	0.92	0.05	0.01
5 dB	6%	62%	89%	58%	59%	5 dB	0.00	0.01	0.90	0.09	0.01
0 dB	7%	45%	79%	44%	37%	0 dB	0.00	0.00	0.90	0.10	0.00
-5 dB	24%	45%	70%	43%	33%	-5 dB	0.00	0.00	0.93	0.07	0.00
N3, best	clean	N1	N2	N3	N4	**N3, 300**	clean	N1	N2	N3	N4
clean	99%	63%	81%	76%	47%	clean	1.00	0.00	0.00	0.00	0.00
20 dB	89%	94%	96%	97%	96%	20 dB	0.02	0.05	0.36	0.51	0.05
15 dB	62%	93%	95%	95%	95%	15 dB	0.00	0.05	0.22	0.67	0.05
10 dB	22%	88%	92%	93%	92%	10 dB	0.00	0.03	0.25	0.70	0.03
5 dB	8%	76%	77%	90%	82%	5 dB	0.00	0.01	0.19	0.79	0.01
0 dB	7%	53%	50%	86%	63%	0 dB	0.00	0.01	0.34	0.66	0.00
-5 dB	35%	59%	57%	78%	63%	-5 dB	0.00	0.00	0.49	0.51	0.00
N4, best	clean	N1	N2	N3	N4	**N4, 300**	clean	N1	N2	N3	N4
clean	99%	64%	79%	77%	45%	clean	1.00	0.00	0.00	0.00	0.00
20 dB	83%	92%	92%	92%	96%	20 dB	0.02	0.13	0.05	0.01	0.80
15 dB	56%	91%	87%	89%	97%	15 dB	0.00	0.17	0.04	0.02	0.77
10 dB	24%	84%	78%	81%	96%	10 dB	0.00	0.13	0.03	0.01	0.82
5 dB	8%	76%	65%	68%	91%	5 dB	0.00	0.11	0.09	0.01	0.79
0 dB	6%	66%	44%	54%	84%	0 dB	0.00	0.15	0.25	0.14	0.46
-5 dB	17%	58%	35%	47%	70%	-5 dB	0.00	0.13	0.37	0.28	0.22

Table A.3: Full confusion matrix for blind acoustic model selection (Aurora 2). The left hand side of the table shows the percentage of utterances of each test set (clean and N1 to N4 of different SNRs) for which the respective set of acoustic models provides the best word accuracy rates. Acoustic models without mismatch usually perform best. In case of noise all acoustic models trained on one of the four noise domains perform almost equally well. The right hand side shows the confusion for blind acoustic model selection and a compact representation of size $k = 300$. Confusion mainly occurs for low SNRs.

	domain specific				is	multi cond.	
	tc	**tn**	**rc**	**rn**	**domain**	**16 GMs**	**64 GMs**
tc	33.81	23.94	11.72	3.09	33.81	30.55	34.82
tn	38.59	52.87	14.56	21.95	52.87	47.44	53.85
rc	21.16	19.32	56.15	36.50	56.15	47.53	54.50
rn	15.31	28.09	24.53	53.98	53.98	48.63	54.89
all	25.52	34.30	23.66	35.76	51.48	45.96	52.12

Table A.4: Baseline phoneme accuracy rates in % for MoveOn evaluation set. The baseline results for domain-specific acoustic models (tc: throat clean, tn: throat noisy, rc: right clean, rn: right noisy) with 16 Gaussian mixtures and multi-conditional acoustic models with 16 and 64 Gaussian mixtures are presented. Multi-conditional acoustic models with 64 mixtures provide the highest average phoneme accuracy rates.

	same domain					average	multi cond.	
SNR	**clean**	**N1**	**N2**	**N3**	**N4**	**domain**	**3 GMs**	**15 GMs**
clean	99.04	—	—	—	—	99.04	98.73	99.45
20dB	—	98.03	98.37	98.18	97.50	98.02	97.96	98.94
15dB	—	97.64	97.76	97.49	97.75	97.66	97.06	98.40
10dB	—	96.47	96.07	96.36	95.71	96.15	95.13	96.97
5dB	—	92.26	90.18	91.53	91.18	91.29	88.39	92.05
0dB	—	77.56	64.87	71.13	75.87	72.36	64.58	73.55
-5dB	—	37.95	25.83	25.41	40.39	32.39	26.68	34.65
all	—	—	—	—	—	83.84	81.22	84.86
20-0dB	—	92.39	89.45	90.94	91.60	91.10	88.63	91.98

Table A.5: Baseline word accuracy rates in % for Aurora 2 evaluation set. The baseline results for domain-specific acoustic models (clean, as well as noise types N1, N2, N3 and N4) with 3 Gaussian mixtures and multi-conditional acoustic models with 3 and 15 Gaussian mixtures are presented. Multi-conditional acoustic models with 64 mixtures provide the highest average word accuracy rates.

	baseline				rQCN			
	tc	**tn**	**rc**	**rn**	**tc**	**tn**	**rc**	**rn**
tc	33.81	23.94	11.72	3.09	34.44	24.63	11.14	3.04
tn	38.59	52.87	14.56	21.95	38.38	52.43	14.51	21.26
rc	21.16	19.32	56.15	36.50	21.19	20.05	56.34	37.15
rn	15.31	28.09	24.53	53.98	14.77	27.86	24.25	53.35
all	25.52	34.30	23.66	35.76	25.29	34.24	23.47	35.22

Table A.6: Full table of phoneme accuracy rates in % for rQCN (MoveOn). The table presents detailed recognition results for rQCN for matched and mismatched conditions of domain-specific acoustic models (tc: throat clean, tn: throat noisy, rc: right clean, rn: right noisy). Phoneme accuracy rates for clean close-talk microphone test data (rc) are slightly improved by rQCN in all cases. For clean throat microphone test data (tc) improvements are shown for both throat microphone acoustic models. In most other cases phoneme accuracy rates slightly decrease for rQCN.

	baseline				rCGN			
	tc	**tn**	**rc**	**rn**	**tc**	**tn**	**rc**	**rn**
tc	33.81	23.94	11.72	3.09	34.00	24.72	11.35	3.89
tn	38.59	52.87	14.56	21.95	38.41	52.49	14.41	21.61
rc	21.16	19.32	56.15	36.50	20.83	19.81	55.96	36.69
rn	15.31	28.09	24.53	53.98	14.52	27.85	23.95	53.02
all	25.52	34.30	23.66	35.76	25.09	34.23	23.29	35.34

Table A.7: Full table of phoneme accuracy rates in % for rCGN (MoveOn). The table presents detailed recognition results for rCGN for matched and mismatched conditions of domain-specific acoustic models (tc: throat clean, tn: throat noisy, rc: right clean, rn: right noisy). Phoneme accuracy rates for clean close-talk microphone test data (rc) and clean throat microphone test data (tc) are slightly improved by rCGN in some cases. In most other cases phoneme accuracy rates slightly decrease when applying rCGN.

Bibliography

[1] T. Winkler, T. Kostoulas, R. Adderley, C. Bonkowski, T. Ganchev, J. Köhler and N. Fakotakis, 'The MoveOn Motorcycle Speech Corpus', *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008.

[2] T. Kostoulas, T. Winkler, T. Ganchev, N. Fakotakis and J. Köhler, 'The MoveOn Database: Motorcycle Environment Speech and Noise Database for Command and Control Applications', Submitted to Springer Journal of Language Resources and Evaluation.

[3] T. Winkler and R. Bardeli,
'An Integrated Approach for a Robust Command and Control Application on the Motorcycle',
Proceedings of the 13th International Conference "Speech and Computer" SPECOM 2009,
2009 464–469.

[4] T. Winkler, S. Pronkine, R. Bardeli and J. Köhler, 'A Study of Throat Microphone Performance in Automatic Speech Recognition on Motorcycles',
Proceedings of the NAG/DAGA 2009 International Conference on Acoustics, 2009.

[5] T. Winkler, 'How Realistic is Artificially Added Noise?', *Proceedings of the 12th Annual Conference of the International Speech Communication Association ISCA (INTERSPEECH)*, International Speech Communication Association, 2011 2605–2608.

[6] D. Stein, T. Winkler and J. Schwenninger, 'Harmonic Distortion in the TETRA Channel and its Impact on Automatic Speech Recognition', *Fortschritte der Akustik - Tagungsband der 38. Deutschen Jahrestagung für Akustik DAGA 2012 in Darmstadt*, Mar. 2012.

[7] D. Stein, T. Winkler, J. Schwenninger and R. Bardeli,
'TETRA Channel Simulation for Automatic Speech Recognition',
Proceedings of the 20th European Signal Processing Conference 2012 (EUSIPCO 2012),
Aug. 2012.

[8] T. Winkler, D. Stein, R. Bardeli, D. Schneider and J. Köhler, 'Potentials for ASR based on Multiple Acoustic Models and Model Selection using Standard Speech Features',
Accepted for Speech Communication, 10. ITG Fachtagung Sprachkommunikation 2012,
Sept. 2012.

[9] K. H. Davis, R. Biddulph and S. Balashek, 'Automatic recognition of spoken digits',
Journal of the Acoustical Society of America 24 (1952) 637–642.

[10] H. F. Olson and H. Belar, 'Phonetic typewriter',
Journal of the Acoustical Society of America 28 (1956) 1072–1081.

[11] T. B. Martin, A. L. Nelson and H. J. Zadell,
'Speech Recognition by Feature-abstraction Techniques', tech. rep.,
Air Force Avionics Laboratory, 1964.

[12] T. K. Vintsyuk, 'Speech discrimination by dynamic programming',
Kibernetika 4.2 (1968) 81–88.

[13] A. Viterbi, 'Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm', *IEEE Transactions on Information Theory* IT-13 (1967) 260–269.

[14] J. G. David Forney, 'The Viterbi Algorithm', *Proceedings of the IEEE*, vol. 61, 3, Mar. 1973 268–278.

[15] F. Itakura, 'Minimum prediction Residual Principle Applied to Speech Recognition', *IEEE Transactions on Acoustics, Speech and Signal Processing* 23 (1975) 169–176.

[16] A. Newell and Carnegie-Mellon University. Computer Science Dept., *Harpy, Production Systems and Human Cognition*, CMU-CS, Carnegie-Mellon University, Department of Computer Science, 1978.

[17] X. Huang, Y. Ariki and M. Jack, *Hidden Markov Models for Speech Recognition*, New York, NY, USA: Columbia University Press, 1990.

[18] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series, PTR Prentice Hall, 1993.

[19] L. R. Bahl, F. Jelinek and R. L. Mercer, 'A maximum likelihood approach to continuous speech recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5 (1983) 179–190.

[20] J. S. Bridle and M. D. Brown, 'An Experimental Automatic Word-Recognition System', tech. rep. 1003, Joint Speech Research Unit, 1974.

[21] P. Mermelstein, 'Distance measure for speech recognition, psychological and instrumental', *Joint Workshop on Pattern Recognition and Artificial Intelligence*, 1976.

[22] S. Furui, 'Speaker-independent isolated word recognition using dynamic features of speech spectrum', *IEEE Transactions on Acoustics, Speech and Signal Processing* 34.1 (Feb. 1986) 52 –59.

[23] C. J. Leggetter and P. C. Woodland, 'Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models', *Computer Speech & Language* 9.2 (Apr. 1995) 171–185.

[24] M. J. F. Gales and S. J. Young, 'Parallel model combination for speech recognition in noise', tech. rep., University of Cambridge: Department of Engineering, 1993.

[25] D. Schneider, J. Schon and S. Eickeler, 'Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System', *Proceedings of the ACM SIGIR Workshop "Searching Spontaneous Conversational Speech" held at SIGIR '08*, ed. by J. Köhler, M. Larson, F. Jong de, W. Kraaij and R. Ordelman, Singapore, July 2008.

[26] Nuance Communications, *Dragon NaturallySpeaking Professional for Social Services Agencies*, White Paper, Nov. 2008, URL: `http://www.nuance.com/`.

[27] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar and B. Strope, 'Google Search by Voice: A Case Study', *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, Springer, 2010 61–90.

[28] B. Raj and R. M. Stern, 'Missing-Feature Approaches in Speech Recognition', *IEEE Signal Processing Magazine* 22.5 (Sept. 2005) 101–116.

[29] J. Chen, J. Benesty, Y. A. Huang and S. Doclo,
'New insights into the noise reduction Wiener filter',
IEEE Transactions on Audio, Speech & Language Processing 14.4 (2006) 1218–1234.

[30] M. L. Seltzer, B. Raj and R. M. Stern,
'Likelihood-Maximizing Beamforming for Robust Hands-Free Speech Recognition',
IEEE Transactions on Speech and Audio Processing 12.5 (Sept. 2004) 489–498.

[31] W. X. Teng, G. Gravier, F. Bimbot and F. Soufflet, 'Speaker adaptation by variable reference model subspace and application to large vocabulary speech recognition', *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, Washington, DC, USA: IEEE Computer Society, 2009 4381–4384.

[32] M. Seltzer, A. Acero and K. Kalgaonkar,
'Acoustic model adaptation via Linear Spline Interpolation for robust speech recognition',
Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010), Mar. 2010 4550 –4553.

[33] S. Furui, '50 Years of Progress in Speech and Speaker Recognition Research',
ECTI Transactions on Computer and Information Technology 1 (Nov. 2005).

[34] B. H. Juang and L. R. Rabiner, 'Elsevier Encyclopedia of Language and Linguistics',
Elsevier, 2005,
chap. Automatic speech recognition - A brief history of the technology development 806–819.

[35] J. Benesty, M. M. Sondhi and Y. Huang, eds., *Springer Handbook of Speech Processing*,
Berlin: Springer, 2008.

[36] E. G. Schukat-Talamazzini,
Automatische Spracherkennung - Grundlagen, statistische Modelle und effiziente Algorithmen,
Künstliche Intelligenz, Vieweg, 1995 I–XI, 1–403.

[37] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland,
The HTK Book Version 3.4, Cambridge University Press, 2006.

[38] L. D. Alsteris and K. K. Paliwal,
'Short-time phase spectrum in speech processing: A review and some experimental results',
Digital Signal Processing 17.3 (May 2007) 578–616.

[39] S. S. Stevens, J. Volkman and E. Newman,
'A scale for the measurement of the psychological magnitude pitch',
Acoustical Society of America 8 (1937) 185–190.

[40] B. P. Bogert, M. J. R. Healy and J. W. Tukey, 'The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking',
Proceedings of the Symposium on Time Series Analysis, 1963 209–243.

[41] I. P. Association, *Handbook of the International Phonetic Association*,
Cambridge University Press, 1999.

[42] Y. Wu, A. Ganapathiraju and J. Picone,
'Report for Baum-Welch Re-estimation of Hidden Markov Model', tech. rep.,
Institute for Signal and Information Processing, Mississippi State University, June 1999.

[43] L. E. Baum, 'An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes', *Inequalities* 1 (1972) 1–8.

[44] H. Schwenk and J.-L. Gauvain,
'Training neural network language models on very large corpora',
Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05,
Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005 201–208.

[45] T. Mikolov, A. Deoras, D. Povey, L. Burget and J. Černocký,
'Strategies for Training Large Scale Neural Network Language Models', *Proceedings of the 2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*,
Hawaii, US: IEEE Signal Processing Society, 2011 196–201.

[46] J. Ming, 'Noise compensation for speech recognition with arbitrary additive noise',
IEEE Transactions on Audio, Speech, and Language Processing 14.3 (May 2006) 833–844.

[47] D. Pearce and H. G. Hirsch, 'The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions', *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*,
Bejing, China, Oct. 2000.

[48] D. Baum, D. Schneider, R. Bardeli, J. Schwenninger, B. Samlowski, T. Winkler and J. Köhler, 'DiSCo - A German Evaluation Corpus for Challenging Problems in the Broadcast Domain', *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, ed. by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner and D. Tapias,
Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[49] B. Raj, V. Parikh and R. Stern,
'The Effects of Background Music on Speech Recognition Accuracy',
Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997) - Volume 2, Washington, DC, USA: IEEE Computer Society, 1997 851.

[50] P. Vanroose, 'Blind Source Separation of Speech and Background Music for Improved Speech Recognition', *Proceedings of the 24th Symposium on Information Theory in the Benelux*, 2003 103–108.

[51] J. R. Hershey, S. J. Rennie, P. A. Olsen and T. T. Kristjansson,
'Super-Human Multi-Talker Speech Recognition: A Graphical Modeling Approach',
Computer Speech & Language In Press, Accepted Manuscript (2009).

[52] S. F. Boll, 'Suppression of Acoustic Noise in Speech Using Spectral Subtraction',
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1979. ICASSP '79, vol. ASSP-27, 2, Apr. 1979 113–120.

[53] M. Berouti, R. Schwartz and J. Makhoul,
'Enhancement of speech corrupted by acoustic noise', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1979)*, vol. 4, Apr. 1979 208 –211.

[54] S. Thomas, S. Ganapathy and H. Hermansky,
'Recognition of Reverberant Speech Using Frequency Domain Linear Prediction',
Signal Processing Letters, IEEE 15 (2008) 681–684.

[55] D. Hirtle, 'Speech Variability: The Biggest Hurdle for Recognition', tech. rep., Faculty of Computer Science, University of New Brunswick, 2004.

[56] M. Benzeghiba, R. de Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris et al., 'Automatic speech recognition and speech variability: A review.', *Speech Communication* 49.10-11 (2007) 763–786.

[57] A. Andreou, T. Kamm and J. Cohen, 'Experiments in Vocal Tract Normalization', *Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[58] L Lee and R. C. Rose, 'Speaker normalization using efficient frequency warping procedures', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996 (ICASSP 1996)*, vol. 1, IEEE, 1996 353–356.

[59] M. Nakamura, K. Iwano and S. Furui, 'Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance', *Computer Speech & Language* 22.2 (Apr. 2008) 171–184.

[60] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie and C. Cox, 'ASR for emotional speech: Clarifying the issues and enhancing performance.', *Neural Networks* 18.4 (2005) 437–444.

[61] B. Schuller, S. Steidl, A. Batliner, F. Schiel and J. Krajewski, 'The INTERSPEECH 2011 Speaker State Challenge', *Proceedings of the 12th Annual Conference of the International Speech Communication Association ISCA (INTERSPEECH)*, Florence, Italy: ISCA, Aug. 2011 3201–3204.

[62] F. Schiel, 'Perception of Alcoholic Intoxication in Speech', *Proceedings of the 12th Annual Conference of the International Speech Communication Association ISCA (INTERSPEECH)*, Florence, Italy: ISCA, Aug. 2011 3281–3284.

[63] Étienne Lombard, 'Le signe de l'élévation de la voix', *Annales des Maladies de L'oreille, du Larynx, du Nez et du Pharynx* 37 (1911) 101–119.

[64] J.-C. Junqua, S. Fincke and K. Field, 'The Lombard effect: a reflex to better communicate with others in noise', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999 (ICASSP '99)*, vol. 4, Mar. 1999 2083–2086 vol.4.

[65] A. Wakao, K. Takeda and F. Itakura, 'Variability of Lombard effects under different noise conditions', *Proceedings of the Fourth International Conference on Spoken Language, 1996 (ICSLP 96)*, vol. 4, Oct. 1996 2009–2012 vol.4.

[66] L. Folk and F. Schiel, 'The Lombard Effect in Spontaneous Dialog Speech', *Proceedings of the 12th Annual Conference of the International Speech Communication Association ISCA (INTERSPEECH)*, Florence, Italy: ISCA, Aug. 2011 2701–2704.

[67] G. Bapineedu, 'Analysis of lombard effect speech and its application in speaker verification for imposter detection', MA thesis: International Institute of Information Technology, 2010.

[68] S.-M. Chi and Y.-H. Oh, 'Lombard effect compensation and noise suppression for noisy Lombard speech recognition', *Proceedings of the Fourth International Conference on Spoken Language, 1996 (ICSLP 96)*, vol. 4, Oct. 1996 2013–2016 vol.4.

[69] S. Pronkine, 'Evaluation von Verfahren zur Aufbereitung von Kehlkopfmikrofonsignalen für eine robuste Spracherkennung', Diploma thesis: University of Bonn, 2009.

[70] H. Bořil and J. H. L. Hansen, 'Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments',
IEEE Transactions on Audio, Speech, and Language Processing 18.6 (Aug. 2010) 1379–1393.

[71] E. R. Geddes and L. W. Lee, 'Auditory Perception of Nonlinear Distortion',
Audio Engineering Society Convention 115, Oct. 2003.

[72] E. R. Geddes and L. W. Lee, 'Auditory Perception of Nonlinear Distortion - Theory',
Audio Engineering Society Convention 115, Oct. 2003.

[73] S. Euler and J. Zinke,
'The Influence of Speech Coding Algorithms on Automatic Speech Recognition', *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1994)*, vol. i, Apr. 1994 I/621–I/624.

[74] H. Saruwatari, Y. Ishikawa, Y. Takahashi, T. Inoue, K. Shikano and K. Kondo,
'Musical Noise Controllable Algorithm of Channelwise Spectral Subtraction and Adaptive Beamforming Based on Higher Order Statistics.',
IEEE Transactions on Audio, Speech & Language Processing 19.6 (2011) 1457–1466.

[75] S. V. Vaseghi, 'Advanced Digital Signal Processing and Noise Reduction', 2nd ed.,
John Wiley & Sons Ltd, 2000, chap. Spectral Subtraction 333–354.

[76] B. S. Atal, 'Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification', *Acoustical Society of America* 55 (1974) 1304–1312.

[77] O. Viikki and K. Laurila, 'Noise robust HMM- based speech recognition using segmental cepstral feature vector normalization' (1997).

[78] S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga,
'Cepstral gain normalization for noise robust speech recognition', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004 (ICASSP '04)*, vol. 1, May 2004 I –209–12 vol.1.

[79] S. Dharanipragada and M. Padmanabhan,
'A Nonlinear Unsupervised Adaptation Technique for Speech Recognition',
Proceedings of the International Conference on Spoken Language Processing, 2000 556–559.

[80] F. Hilger, S. Molau and H. Ney,
'Quantile based histogram equalization for online applications',
Proceedings of the 7th International Conference on Spoken Language Processing, 2002 237–240.

[81] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*, Cambridge, MA, USA: MIT Press, 1949.

[82] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*,
John Wiley & Sons Ltd, 2000.

[83] P. Zhan and A. Waibel,
'Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition',
tech. rep., School of Computer Science, Carnegie Mellon University, 1997.

[84] D. R. Sanand and S. Umesh,
'VTLN Using Analytically Determined Linear-Transformation on Conventional MFCC',
IEEE Transactions on Audio, Speech, and Language Processing 20.5 (July 2012) 1573–1583.

[85] A. Acero and X. Huang, 'Augmented Cepstral Normalization for Robust Speech Recognition',
Proceedings of the IEEE Workshop on Automatic Speech Recognition, 1995.

[86] *ETSI ES 202 050 V1.1.5*, http://www.etsi.org,
European Telecommunications Standards Institute (ETSI), Jan. 2007.

[87] J.-Y. Li, B. Liu, R.-H. Wang and L.-R. Dai,
'A Complexity Reduction of ETSI Advanced Front-End for DSR', *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, vol. 1, 2004 61–64.

[88] J. C. Segura, M. Benítez, A. de la Torre and A. J. Rubio, 'Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR',
Proceedings of the International Conference on Spoken Language Processing (ICSLP-2002), Denver, CO, USA, 2002 225–228.

[89] J.-J. Gauvain and C.-H. Lee, 'Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains',
IEEE Transactions on Speech and Audio Processing 2.2 (Apr. 1994) 291–298.

[90] P. Červa and J. Nouza,
'MAP Based Speaker Adaptation in Very Large Vocabulary Speech Recognition of Czech',
Radioengineering 13.3 (2004) 42–46.

[91] M. Gales and P. Woodland, 'Mean and Variance Adaptation within the MLLR Framework',
Computer Speech & Language 10 (1996) 249–264.

[92] M. L. Seltzer and A. Acero,
'Separating Speaker and Environmental Variability Using Factored Transforms',
Proceedings of the 12th Annual Conference of the International Speech Communication Association ISCA (INTERSPEECH 2011), Florence, Italy: ISCA, Aug. 2011 1097–1100.

[93] H.-G. Hirsch and A. Kitzig, 'Improving the robustness with multiple sets of HMMs',
Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH 2009), 2009 564–567.

[94] H. Xu, Z.-H. Tan, P. Dalsgaard and B. Lindberg, 'Robust Speech Recognition Based on Noise and SNR Classification - A Multiple-Model Framework', *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH 2005)*, 2005 977–980.

[95] Y. Gao, M. Padmanabhan and M. Picheny,
'Speaker Adaptation Based on Pre-Clustering Training Speakers', *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-1997)*, Rhodes, Greece, Sept. 1997 2091–2094.

[96] Z. Zhang, T. Sugimura and S. Furui, 'A tree-structured clustering method integrating noise and SNR for piecewise linear-transformation-based noise adaptation', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004 (ICASSP '04)*, vol. 1, May 2004 I –981–4 vol.1.

[97] N. Thatphithakkul, B. Kruatrachue, C. Wutiwiwatchai, S. Marukatat and V. Boonpiam, 'Tree-structured model selection and simulated-data adaptation for environmental and speaker robust speech recognition', *Proceedings of the International Symposium on Communications and Information Technologies, 2007 (ISCIT '07)*, 2007 1570–1574.

[98] D. Becerril, O. Saz, C. Vaquero, A. Ortega and E. Lleida, 'Speaker Tree Generation for Model Selection in Automatic Speech Recognition', *Proceedings of FALA 2010 VI Jornada en Tecnología del Habla and II Iberian SL Tech Workshop*, 2010.

[99] S. S. Chen and P. S. Gopalakrishnan, 'Clustering via the Bayesian information criterion with applications in speech recognition', *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1998 (ICASSP '98)*, vol. 2, 1998 645–648.

[100] O. Saz, E. Lleida, C. Vaquero and W. R. Rodríguez, 'The Alborada-I3A Corpus of Disordered Speech.', *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, ed. by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias, European Language Resources Association, 2010.

[101] V. I. Levenshtein, 'Binary codes capable of correcting deletions, insertions and reversals', *Soviet Physics-Doklady* 10 (1966) 707–710.

[102] R. G. Leonard and G. Doddington, *TIDIGITS*, 1993, URL: `http://www.ldc.upenn.edu/`.

[103] *ITU recommendation G.712: Transmission performance characteristics of pulse code modulation channels*, International Telecommunication Union (ITU), Nov. 1996.

[104] ETSI, *ETSI ETS 300 395-2 ed.2: Terrestrial Trunked Radio (TETRA); Speech codec for full-rate traffic channel; Part 2: TETRA codec*, http://www.etsi.org, European Telecommunication Standard Institute (ETSI), Feb. 1998.

[105] E. Kalapanidas, C. Davarakis, M. Nani, T. Winkler, T. Ganchev, O. Kocsis, N. Fakotakis, A. Handzlik, G. Swiecanski, A. Badii et al., 'MoveON: A Multimodal Information Management Application for Police Motorcyclists', *System Demonstrations of the 18th European Conference on Artificial Intelligence*, Patras, Greece, July 2008.

[106] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura and Y. Inagaki, 'Construction of speech corpus in moving car environment', *Proceedings of the Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000*, Beijing, China, Oct. 2000 362–365.

[107] Y.-J. Lee, B.-W. Kim, Y.-I. Kim, D.-L. Choi, K.-H. Lee and Y. Um, 'Creation and Assessment of Korean Speech and Noise DB in Car Environment', *Proceedings of the Fourth International Conference On Language Resources and Evaluation (LREC 2008)*, 2004 1403–1406.

[108] B. Lee, M. Hasegawa-johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. Huang, 'AVICAR: Audio-Visual Speech Corpus in a Car Environment', *Proceedings of the 8th International Conference on Spoken Language Processing, ICSLP 2004 / INTERSPEECH 2004*, Jeju Island, Korea, Oct. 2004 2489–2492.

[109] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler and J. Allen, 'SPEECHDAT-CAR. A Large Speech Database For Automotive Environments', *Proceedings of the 2nd International Conference on Language Resources and Evaluation, (LREC 2000)*, Athens, Greece, 2000.

[110] M. Kaiser, H. Mögele and F. Schiel, 'Bikers Accessing the Web: The SmartWeb Motorbike Corpus', *Proceedings of the 5th International Conference on Language Resources and Evaluation, (LREC 2006)*, Genova, Italy: ELRA, May 2006 1628–1631.

[111] H. van den Heuvel, L. Boves, A. Moreno, M. Omologo, G. Richard and E. Sanders, 'Annotation in the SpeechDat Projects', *International Journal of Speech Technology* 4.2 (2001) 127–143.

[112] F. Schiel and C. Draxler, *Production and Validation of Speech Corpora*, Books on Demand GmbH, 2003.

[113] P. Boersma and D. Weenink, 'PRAAT, a system for doing phonetics by computer', *Glot International* 5.9/10 (2001) 341–345.

[114] C. M. Whissell, 'The dictionary of affect in language', *Emotion Theory Research and Experience*, ed. by R. Plutchik and H. Kellerman, vol. 4, Academic Press, 1989 113–131.

[115] J. Wells, *Standards, Assessment, and Methods: Phonetic Alphabets*, University College, London, 1997.

[116] S. J. Wheatley and S. R. Ascham, 'SpeechDat English database for the fixed telephone network', tech. rep., 1998.

[117] D. Schneider, J. Schon and S. Eickeler, 'Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System', *Proceedings of the ACM SIGIR Workshop "Searching Spontaneous Conversational Speech" held at SIGIR '08*, ed. by J. Köhler, M. Larson, F. Jong de, W. Kraaij and R. Ordelman, Singapore, July 2008.

[118] Z. Valsan and M. Emele, 'Thematic text clustering for domain specific language model adaptation', *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2003 (ASRU 2003)*, 2003 513–518.

[119] W. Kim, 'Language Model Adaptation for Automatic Speech Recognition and Statistical Machine Translation', PhD thesis: The Johns Hopkins University, Oct. 2004.

[120] D. Schneider, T. Winkler, J. Löffler and J. Schon, 'Robust Audio Indexing and Keyword Retrieval Optimized for the Rescue Operation Domain', *Mobile Response, First International Workshop on Mobile Information Technology for Emergency Response, Mobile Response 2007*, ed. by J. Löffler and M. Klann, Lecture Notes in Computer Science, Sankt Augustin, Germany: Springer, Feb. 2007 135–142.

[121] *VoxForge Acoustic Models*, Acoustic Models, version 0.1.1-build726, URL: `http://www.repository.voxforge1.org/downloads/Main/Tags/Releases`.

[122] D. Stein and B. Usabaev,
'Automatic Speech Recognition on a Firefighter TETRA Broadcast Channel', *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, ed. by N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis,
Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

[123] J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P.-A. Breton, V. Clot, R. Gemello, M. Matassoni and P. Maragos, *The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication*, 2007, URL: http://www.hiwire.org/.

[124] M. Graciarena, H. Franco, K. Sonmez and H. Bratt,
'Combining standard and throat microphones for robust speech recognition',
Signal Processing Letters, IEEE 10.3 (Mar. 2003) 72–74.

[125] S. Dupont and C. Ris, 'Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise', *Proceedings of the Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction)*,
Norwich, UK, Aug. 2004.

[126] S.-C. Jou, T. Schultz and A. Waibel,
'Adaptation for Soft Whisper Recognition Using a Throat Microphone', *Proceedings of the International Conference on Speech and Language Processing, ICSLP 2004*,
Jeju Island, Korea, Oct. 2004.

[127] A. Shahina and B. Yegnanarayana, 'Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach',
EURASIP Journal on Advances in Signal Processing 2007 (2007).

[128] X. Huang, A. Acero and H.-W. Hon,
Spoken Language Processing: A Guide To Theory, Algorithm, And System Development,
Prentice Hall PTR, 2001.

[129] C. Slump, T. Simons and K.A.Verweij, 'On the Objective Speech Quality of TETRA',
Proceedings of the Anual workshop on Circuits, Systems and Signal Processing,
Mierlo, the Netherlands, Nov. 1999 421–429.

[130] M. Steppler, 'Leistungsbewertung von TETRA-Mobilfunksystemen durch Analyse und Emulation ihrer Protokolle', PhD thesis: RWTH Aachen University, July 2002.

[131] A. Preti, B. Ravera, F. Capman and J.-F. Bonastre,
'An Application Constrained Front End for Speaker Verification',
Proceedings of the 16th European Signal Processing (EUSIPCO),
Lausanne, Switzerland, Aug. 2008.

[132] D. A. Reynolds, T. F. Quatieri and R. B. Dunn,
'Speaker Verification Using Adapted Gaussian Mixture Models', *Digital Signal Processing*, vol. 10, 2000 19–41.

List of Figures

List of Tables

Abbreviations and Acronyms

ACELP	Algebraic Code-Excited Linear Prediction
AM (Corpus)	Audio Mining (Corpus); only in context with corpus
AM	Acoustic Models
AMR	Adaptive Multi Rate
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BAMS	Blind Acoustic Model Selection
BEEP	British English Example Pronunciation
BIC	Bayesian Information Criterion
BMW	Bayrische Motoren Werke
BNF	Backus-Naur Form
c&c	Command and Control
CDF	Cumulative Distribution Function
CELP	Code-Excited Linear Prediction
CGN	Cepstral Gain Normalisation
CMN	Cepstral Mean Normalisation
CVN	Cepstral Variance Normalisation
CMLLR	Constrained Maximum Likelihood Linear Regression
dB	Decibel
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
ELDA	Evaluations and Language resources Distribution Agency
EM	Expectation Maximisation
ETSI	European Telecommunications Standards Institute
FE	Feature Extraction
GM	Gaussian Mixture
GMM	Gaussian Mixture Model
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HMM	Hidden Markov Model
HIWIRE	Human Input that Works In Real Environments
HTK	Hidden Markov Model Toolkit
Hz	Hertz
IAIS	Intelligente Analyse und Informationssysteme
IST	Information Society Technologies
LM	Language Model
LPC	Linear Predictive Coding
LSP	Line Spectral Pair

LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum A Posteriori
MIRS	Motorola Integrated Radio System
MIT	Massachusetts Institute of Technology
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MoveOn	Multi-modal and Multi-sensor Zero-distraction Interaction Interface for Two Wheel Vehicles on the Move
MFCC	mel-frequency cepstral coefficient
NIST	National Institute of Standards and Technology
PAR	Phoneme Accuracy Rate
PER	Phoneme Error Rate
PMC	Parallel Model Combination
QCN	Quantile Based Cepstral Dynamics Normalisation
rCGN	relative Cepstral Gain Normalisation
rQCN	relative Quantile Based Cepstral Dynamics Normalisation
SAMPA	Speech Assessment Methods Phonetic Alphabet
SNR	Signal-to-noise Ratio
SoX	Sound eXchange
STFT	Short-time Fourier Transform
TETRA	Terrestrial Trunked Radio
THD	Total Harmonic Distortion
TV	Television
UK	United Kingdom
VAD	Voice Activity Detection
VTLN	Vocal Tract Length Normalisation
VQ	Vector Quantisation
WAR	Word Accuracy Rate
WER	Word Error Rate